CFA Institute

CFA INSTITUTE INVESTMENT SERIES

ECONOMICS FOR INVESTMENT DECISION MAKERS

Micro, Macro, and International Economics

Christopher D. Piros, CFA • Jerald E. Pinto, CFA

Foreword by Larry Harris, PhD, CFA, USC Marshall School of Business

ECONOMICS FOR INVESTMENT DECISION MAKERS

CFA Institute is the premier association for investment professionals around the world, with over 117,000 members in 134 countries. Since 1963 the organization has developed and administered the renowned Chartered Financial Analyst® Program. With a rich history of leading the investment profession, CFA Institute has set the highest standards in ethics, education, and professional excellence within the global investment community and is the foremost authority on investment profession conduct and practice.

Each book in the CFA Institute Investment Series is geared toward industry practitioners along with graduate-level finance students and covers the most important topics in the industry. The authors of these cutting-edge books are themselves industry professionals and academics and bring their wealth of knowledge and expertise to this series.

ECONOMICS FOR INVESTMENT DECISION MAKERS

Micro, Macro, and International Economics

Christopher D. Piros, CFA
Jerald E. Pinto, CFA

Cover Design: Leiva-Sposato
Cover Image: © Maciej Noskowski / iStockphoto

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Economics for investment decision makers : micro, macro, and international economics / Christopher D. Piros and Jerald E. Pinto, editors.
p. cm. — (CFA institute investment series)
Includes bibliographical references and index.
ISBN 978-1-118-10536-8 (cloth); ISBN 978-1-118-41880-2 (ebk);
ISBN 978-1-118-53316-1 (ebk); ISBN 978-1-118-41624-2 (ebk)
1. Supply and demand. 2. Microeconomics. 3. Macroeconomics. 4. Investments.
I. Piros, Christopher Dixon. II. Pinto, Jerald E.
HB171.5.E3356 2013
330—dc23

2012034395

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

FOREWORD

The opportunity to learn economics from a book sponsored by CFA Institute is a special privilege.

Most economics textbooks are written by one or two authors who decide what subjects should appear in them. The results invariably reflect the values and life experiences of the authors, who usually are academics. Academic authors have written many truly excellent economics textbooks on their own, but these sometimes do not speak to the needs of learners whose interest in economics stems from a desire to understand and solve practical economic problems that they regularly encounter, or expect to encounter, in their working lives.

In contrast, highly respected practicing financial analysts and senior academic economists worked together to select the topics that appear in this book. The topics were chosen from the body of knowledge that the CFA Institute Education Advisory Committee identified as topics that CFA Program candidates need to learn for earning the well-regarded Chartered Financial Analyst (CFA) designation. The Candidate Body of Knowledge™ consists of the knowledge, skills, and abilities that are necessary to analyze and solve common practical problems that arise in investing, valuing investments or companies, and managing portfolios.

This volume is an edited compilation of readings on economics from the CFA Program curriculum. The chapters are written by highly regarded economists and practitioners who were asked to present the topics in a way that is readily accessible to everyone. The authors were chosen by CFA Institute for the depth of their understanding of the topics assigned to them and also for their proven ability to effectively teach the topics.

The readings in the CFA Program curriculum always start with a set of learning outcome statements (LOS) that briefly and clearly state what candidates should know after completing the reading. The editors have included the LOS associated with each topic covered in this text to assist you with your learning. If you can confidently and honestly say that you understand the knowledge described in the various learning outcome statements, you will have mastered the knowledge presented in this book.

The CFA Program curriculum also includes practice problems at the end of every reading. These problems allow CFA Program candidates to practice solving practical problems and to measure their learning progress. The editors have included many of these problems (and some others as well) in a workbook to help you learn the material.

For students interested in possibly earning the CFA charter, this book is an excellent introduction to what would be in the CFA Program curriculum. If that is not your objective, don't fret. The topics presented in this volume are of universal interest, and it provides an excellent introduction to economics for all readers.

WHY ECONOMICS?

Economics is the study of how people make decisions when faced with scarce resources. Decisions amenable to economic analysis include business decisions as well as personal decisions. Scarce resources may be tangible things, such as iron or wool; financial assets, such as money or bonds; or intangibles, such as personal time or emotional energy. The decisions that people make influence the actions that they take. Accordingly, understanding economics will help you understand human behavior.

Examples of human behaviors that economics can help you understand include how much companies produce (theory of the firm), why people buy certain products (consumer choice theory), and even how many children families choose to have (demand theory).

THEORY VERSUS PRACTICE

Although the economic topics have been selected for their relevance to practitioners, plenty of economic theory appears in this book. The economic theory is included because it will help you understand economic problems.

This book is full of the application of economics to problems, but you cannot understand the application without understanding the economic theory behind the application. Theory and practice are not antithetical to each other in economics. A thorough understanding of practical problems requires an in-depth understanding of the underlying theory. Accordingly, you must learn economic theory to acquire the skills to understand, analyze, and solve practical problems that interest you.

Some economists have developed theories that do not (yet) have obvious practical applications. Be assured that this book does not present such theories. All theory presented in this book is either necessary to understand practical problems or necessary to understand other theories needed to understand practical problems.

A PRACTICAL EXAMPLE

First, consider some information that you may already know about the oil refining business. The profits that refiners make depend critically on the difference between the price of the crude oil that they buy and the summed prices of the refined products they produce from the oil and then sell. As an aside, this difference is called the "crack spread" because refiners often use a process called "cracking" to split such long-chain hydrocarbons as heavy oil to produce such lighter short-chain products as gasoline.

Now, consider some economic history. Several years ago, before the recent financial crisis occurred, the world economy was growing quickly and the demand for petroleum products, such as gasoline, diesel fuel, and jet fuel, exceeded the capacity of refiners to produce them. As a result, the prices of these fuels rose substantially. The higher product prices discouraged some users from consuming them, so aggregate demand declined and became equal to the capacity of the refiners.

For several years, the high product prices substantially increased the crack spreads and thus refiner profits. Some investors saw these high profits and assumed that they would

continue. So, they bid up the stock prices of the refiners. For various reasons not important to this discussion, the price of crude oil later rose.

Here is the practical question: What effect do you expect the higher price of crude oil had on the value of the common stocks of the refiners? Would you have bought or sold the refiners' common stocks?

When learning economics, putting yourself into the problem is always helpful. When you read this book and see an example, always ask yourself, "What would I do? How would I solve this problem?"

Many investors apparently thought that the higher crude oil prices would further increase the refiners' profits because the refiners sold oil products. Thus, they further bid up the stock prices of oil refiners. Those investors did not understand the economic topics presented in this book.

In fact, the refiners passed on the higher crude oil prices by raising prices for their refined products. Those that did not raise prices would have gone out of business. These higher end-user product prices decreased product demand to the point that the refiners were no longer operating at capacity. The crack spread then fell substantially because the refiners had excess capacity. The common stock of the refiners dropped when investors finally understood the true implications of the higher crude oil prices for the refiners.

If you bought the refiners' common stocks without understanding the underlying economics, you would have lost money, and perhaps also the ability to retire in greater comfort and with greater security. If you sold them short, which means borrowing and selling a security you do not own with the expectation that you will be able to buy it later at a lower price, you would have made money and, with it, the power to command more resources in the economy.

The difference between winners and the losers in the stock market is an in-depth understanding of the underlying economics informed by a trivial understanding of what refiners do. When the refiners were operating at capacity, the refiner's total capacity was a choke point in the supply chain between the wellhead and the consumer—refining capacity then was the most important scarce resource. You will learn in this book that scarce resources earn exceptional profits when they are in high demand. After the price of crude oil rose, the choke point in the supply chain moved from the refiners to the wellhead. Oil producers profited but not the oil refiners.

Studying this book carefully will help you think through problems such as these. This refinery problem touched on several economic topics that appear in this book: supply and demand, derived factor demands, competitive market equilibrium, supply chain dynamics, and the origins of economic rents.

THE ROLE OF THE ABSTRACT

Many people criticize economics for being too theoretical or for drawing conclusions from simplistic assumptions that do not reflect real-world realities. In short, they see economics as being too abstract.

Don't be put off by the use of the abstract in economics. The abstract is a natural result of the desire to identify the most important characteristics of a problem. Understanding problems is much easier when the problem has been reduced to its essentials. With that basis of understanding, you then can add back characteristics of the problem that you initially

assumed away and ask whether or how the result changes. Your job as a student is to continually consider whether essential characteristics of a problem have been assumed away.

Your studies will be most productive when you take a critical view of your lessons. The more you challenge this text and the instructors who assigned it to you, the more you will come to appreciate the value of economic thought.

You may have heard that economists often disagree with each other. That is far less true than it seems. Disagreements make news whereas agreements are not as interesting. Essentially, all reputable economists agree with the ideas presented in this book. But as I noted in the previous paragraph, you should challenge everything you learn.

The disagreements among economists generally are not about their theories but about their personal values. Everyone is entitled to their opinions about what should be. For example, should the U.S. government impose quotas that limit the importing of sugar to protect U.S. farmers who grow corn to process into corn sweetener? Economics cannot answer this question, but it can tell you what the implications are of various policies. For example, imported sugar quotas are the reason why U.S. consumers drink Coca-Cola sweetened with corn sweetener, whereas everyone in the rest of the world drinks Coca-Cola sweetened with sugar. Economists may disagree about whether the government should limit the importation of sugar into the United States, but they all agree on the effects of the policy.

The ideas in this book are very powerful. When you understand them well, you will have the power to apply them throughout your life.

Good luck with your studies!

Larry Harris, PhD, CFA
Fred V. Keenan Chair in Finance
USC Marshall School of Business

ACKNOWLEDGMENTS

We would like to thank the many individuals who played a role in the conception and production of this book. In addition to the authors, these include the following: Robert E. Lamy, CFA; Wendy L. Pirie, CFA; Barbara S. Petitt, CFA; Christopher B. Wiese, CFA; Lamees Al Baharna, CFA; Gary L. Arbogast, CFA; Evan L. Ashcraft, CFA; Sridhar Balakrishna, CFA; Philippe Bernard, CFA; John P. Calverley; Bolong Cao, CFA; Biharilal Laxman Deora, CFA; Jane Farris, CFA; Martha E. Freitag, CFA; John M. Gale; Osman Ghani, CFA; Muhammad J. Iqbal, CFA; William H. Jacobson, CFA; Bryan K. Jordan, CFA; Asjeet S. Lamba, CFA; David Landis, CFA; Konstantinos G. Leonida, CFA; Jay A. Moore, CFA; Murli Rajan, CFA; Raymond D. Rath, CFA; Victoria J. Rati, CFA; Rodrigo F. Ribeiro, CFA; G. D. Rothenburg, CFA; Sanjiv Sabherwal; Joseph D. Shaw, CFA; Sandeep Singh, CFA; Frank E. Smudde, CFA; Zhiyi Song, CFA; Peter C. Stimes, CFA; Oscar Varela, CFA; Lavone Whitmer, CFA; Stephen E. Wilcox, CFA; and Mark E. Wohar.

Christopher D. Piros, CFA, and Jerald E. Pinto, CFA, oversaw development and editing of the book. The Editorial Services group at CFA Institute provided extraordinary support of the book's copyediting needs. Wanda Lauziere of CFA Institute expertly served as project manager for the book's production.

We thank all for their excellent and detailed work.

ABOUT THE CFA INSTITUTE INVESTMENT SERIES

CFA Institute is pleased to provide you with the CFA Institute Investment Series, which covers major areas in the field of investments. We provide this best-in-class series for the same reason we have been chartering investment professionals for more than 45 years: to lead the investment profession globally by setting the highest standards of ethics, education, and professional excellence.

The books in the CFA Institute Investment Series contain practical, globally relevant material. They are intended both for those contemplating entry into the extremely competitive field of investment management as well as for those seeking a means of keeping their knowledge fresh and up to date. This series was designed to be user friendly and highly relevant.

We hope you find this series helpful in your efforts to grow your investment knowledge, whether you are a relatively new entrant or an experienced veteran ethically bound to keep up to date in the ever-changing market environment. As a long-term, committed participant in the investment profession and a not-for-profit global membership association, CFA Institute is pleased to provide you with this opportunity.

THE TEXTS

One of the most prominent texts over the years in the investment management industry has been Maginn and Tuttle's *Managing Investment Portfolios: A Dynamic Process*. The third edition updates key concepts from the 1990 second edition. Some of the more experienced members of our community own the prior two editions and will add the third edition to their libraries. Not only does this seminal work take the concepts from the other readings and put them in a portfolio context, but it also updates the concepts of alternative investments, performance presentation standards, portfolio execution, and, very importantly, individual investor portfolio management. Focusing attention away from institutional portfolios and toward the individual investor makes this edition an important and timely work.

Quantitative Investment Analysis focuses on some key tools that are needed by today's professional investor. In addition to classic time value of money, discounted cash flow applications, and probability material, there are two aspects that can be of value over traditional thinking.

The first involves the chapters dealing with correlation and regression that ultimately figure into the formation of hypotheses for purposes of testing. This gets to a critical skill that challenges many professionals: the ability to distinguish useful information from the overwhelming quantity of available data. For most investment researchers and managers, their analysis is not solely the result of newly created data and tests that they perform. Rather, they

synthesize and analyze primary research done by others. Without a rigorous manner by which to explore research, you cannot understand good research or have a basis on which to evaluate less rigorous research.

Second, the final chapter of *Quantitative Investment Analysis* covers portfolio concepts and takes the reader beyond the traditional capital asset pricing model (CAPM) type of tools and into the more practical world of multifactor models and arbitrage pricing theory.

Fixed Income Analysis has been at the forefront of new concepts in recent years, and this particular text offers some of the most recent material for the seasoned professional who is not a fixed-income specialist. The application of option and derivative technology to the once-staid province of fixed income has helped contribute to an explosion of thought in this area. Professionals have been challenged to stay up to speed with credit derivatives, swaptions, collateralized mortgage securities, mortgage-backed securities, and other vehicles, and this explosion of products has strained the world's financial markets and tested central banks to provide sufficient oversight. Armed with a thorough grasp of the new exposures, the professional investor is much better able to anticipate and understand the challenges our central bankers and markets face.

International Financial Statement Analysis is designed to address the ever-increasing need for investment professionals and students to think about financial statement analysis from a global perspective. The text is a practically oriented introduction to financial statement analysis that is distinguished by its combination of a true international orientation, a structured presentation style, and abundant illustrations and tools covering concepts as they are introduced in the text. The authors cover this discipline comprehensively and with an eye to ensuring the reader's success at all levels in the complex world of financial statement analysis.

Equity Asset Valuation is a particularly cogent and important resource for anyone involved in estimating the value of securities and understanding security pricing. A well-informed professional knows that the common forms of equity valuation—dividend discount modeling, free cash flow modeling, price/earnings modeling, and residual income modeling—can all be reconciled with one another under certain assumptions. With a deep understanding of the underlying assumptions, the professional investor can better understand what other investors assume when calculating their valuation estimates. This text has a global orientation, including emerging markets. The second edition provides new coverage of private company valuation and expanded coverage of required rate of return estimation.

Investments: Principles of Portfolio and Equity Analysis provides an accessible yet rigorous introduction to portfolio and equity analysis. Portfolio planning and portfolio management are presented within a context of up-to-date, global coverage of security markets, trading, and market-related concepts and products. The essentials of equity analysis and valuation are explained in detail and profusely illustrated. The book includes coverage of practitioner-important but often neglected topics, such as industry analysis. Throughout, the focus is on the practical application of key concepts with examples drawn from both emerging and developed markets. Each chapter affords the reader many opportunities to self-check his or her understanding of topics. In contrast to other texts, the chapters are collaborations of respected senior investment practitioners and leading business school faculty from around the globe. By virtue of its well-rounded, expert, and global perspectives, the book should be of interest to anyone who is looking for an introduction to portfolio and equity analysis.

The New Wealth Management: The Financial Advisor's Guide to Managing and Investing Client Assets is an updated version of Harold Evensky's mainstay reference guide for wealth managers. Harold Evensky, Stephen Horan, and Thomas Robinson have updated the core text of the 1997 first edition and added an abundance of new material to fully reflect today's

investment challenges. The text provides authoritative coverage across the full spectrum of wealth management and serves as a comprehensive guide for financial advisors. The book expertly blends investment theory and real-world applications and is written in the same thorough but highly accessible style as the first edition.

Corporate Finance: A Practical Approach is a solid foundation for those looking to achieve lasting business growth. In today's competitive business environment, companies must find innovative ways to enable rapid and sustainable growth. This text equips readers with the foundational knowledge and tools for making smart business decisions and formulating strategies to maximize company value. It covers everything from managing relationships between stakeholders to evaluating merger and acquisition bids, as well as the companies behind them. The second edition of the book preserves the hallmark conciseness of the first edition while expanding coverage of dividend policy, share repurchases, and capital structure. Through extensive use of real-world examples, readers will gain critical perspective into interpreting corporate financial data, evaluating projects, and allocating funds in ways that increase corporate value. Readers will gain insights into the tools and strategies used in modern corporate financial management.

ECONOMICS FOR INVESTMENT DECISION MAKERS

CHAPTER 1

DEMAND AND SUPPLY ANALYSIS: INTRODUCTION

Richard V. Eastin

Gary L. Arbogast, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Distinguish among types of markets.
- Explain the principles of demand and supply.
- Describe causes of shifts in and movements along demand and supply curves.
- Describe the process of aggregating demand and supply curves, the concept of equilibrium, and mechanisms by which markets achieve equilibrium.
- Distinguish between stable and unstable equilibria and identify instances of such equilibria.
- Calculate and interpret individual and aggregate demand and inverse demand and supply functions, and interpret individual and aggregate demand and supply curves.
- Calculate and interpret the amount of excess demand or excess supply associated with a nonequilibrium price.
- Describe the types of auctions and calculate the winning price(s) of an auction.
- Calculate and interpret consumer surplus, producer surplus, and total surplus.
- Analyze the effects of government regulation and intervention on demand and supply.
- Forecast the effect of the introduction and the removal of a market interference (e.g., a price floor or ceiling) on price and quantity.
- Calculate and interpret price, income, and cross-price elasticities of demand, and describe factors that affect each measure.

1. INTRODUCTION

In a general sense, **economics** is the study of production, distribution, and consumption and can be divided into two broad areas of study: macroeconomics and microeconomics. **Macroeconomics** deals with aggregate economic quantities, such as national output and national income. Macroeconomics has its roots in **microeconomics**, which deals with markets and decision making of individual economic units, including consumers and businesses. Microeconomics is a logical starting point for the study of economics.

This chapter focuses on a fundamental subject in microeconomics: demand and supply analysis. **Demand and supply analysis** is the study of how buyers and sellers interact to determine transaction prices and quantities. As we will see, prices simultaneously reflect both the value to the buyer of the next (or marginal) unit and the cost to the seller of that unit. In private enterprise market economies, which are the chief concern of investment analysts, demand and supply analysis encompasses the most basic set of microeconomic tools.

Traditionally, microeconomics classifies private economic units into two groups: consumers (or households) and firms. These two groups give rise, respectively, to the theory of the consumer and theory of the firm as two branches of study. The **theory of the consumer** deals with **consumption** (the demand for goods and services) by utility-maximizing individuals (i.e., individuals who make decisions that maximize the satisfaction received from present and future consumption). The **theory of the firm** deals with the supply of goods and services by profit-maximizing firms. The theory of the consumer and the theory of the firm are important because they help us understand the foundations of demand and supply. Subsequent chapters will focus on the theory of the consumer and the theory of the firm.

Investment analysts, particularly equity and credit analysts, must regularly analyze products and services—their costs, prices, possible substitutes, and complements—to reach conclusions about a company's profitability and business risk (risk relating to operating profits). Furthermore, unless the analyst has a sound understanding of the demand and supply model of markets, he or she cannot hope to forecast how external events—such as a shift in consumer tastes or changes in taxes and subsidies or other intervention in markets—will influence a firm's revenue, earnings, and cash flows.

Having grasped the tools and concepts presented in this chapter, the reader should also be able to understand many important economic relationships and facts and be able to answer questions such as:

- Why do consumers usually buy more when the price falls? Is it irrational to violate this law of demand?
- What are appropriate measures of how sensitive the quantity demanded or supplied is to changes in price, income, and prices of other goods? What affects those sensitivities?
- If a firm lowers its price, will its total revenue also fall? Are there conditions under which revenue might rise as price falls, and, if so, what are those conditions? Why might this occur?
- What is an appropriate measure of the total value consumers or producers receive from the opportunity to buy and sell goods and services in a free market? How might government intervention reduce that value, and what is an appropriate measure of that loss?
- What tools are available that help us frame the trade-offs that consumers and investors face as they must give up one opportunity to pursue another?

- Is it reasonable to expect markets to converge to an equilibrium price? What are the conditions that would make that equilibrium stable or unstable in response to external shocks?
- How do different types of auctions affect price discovery?

This chapter is organized as follows. Section 2 explains how economists classify markets. Section 3 covers the basic principles and concepts of demand and supply analysis of markets. Section 4 introduces measures of sensitivity of demand to changes in prices and income. A summary and a set of practice problems conclude the chapter.

2. TYPES OF MARKETS

Analysts must understand the demand and supply model of markets because all firms buy and sell in markets. Investment analysts need at least a basic understanding of those markets and the demand and supply model that provides a framework for analyzing them.

Markets are broadly classified as factor markets or goods markets. **Factor markets** are markets for the purchase and sale of factors of production. In capitalist private enterprise economies, households own the **factors of production** (the land, labor, physical capital, and materials used in production). **Goods markets** are markets for the output of production. From an economics perspective, firms, which ultimately are owned by individuals either singly or in some corporate form, are organizations that buy the services of those factors. Firms then transform those services into intermediate or final goods and services. (**Intermediate goods and services** are those purchased for use as inputs to produce other goods and services, whereas final goods and services are in the final form purchased by households.) These two types of interaction between the household sector and the firm sector—those related to goods and those related to services—take place in factor markets and goods markets, respectively.

In the factor market for labor, households are sellers and firms are buyers. In goods markets, firms are sellers and both households and firms are buyers. For example, firms are buyers of capital goods (such as equipment) and intermediate goods, while households are buyers of a variety of durable and nondurable goods. Generally, market interactions are *voluntary*. Firms offer their products for sale when they believe the payment they will receive exceeds their cost of production. Households are willing to purchase goods and services when the value they expect to receive from them exceeds the payment necessary to acquire them. Whenever the perceived value of a good exceeds the expected cost to produce it, a potential trade can take place. This fact may seem obvious, but it is fundamental to our understanding of markets. If a buyer values something more than a seller, not only is there an opportunity for an exchange, but that exchange will make *both* parties better off.

In one type of factor market, called **labor markets**, households offer to sell their labor services when the payment they expect to receive exceeds the value of the leisure time they must forgo. In contrast, firms hire workers when they judge that the value of the productivity of workers is greater than the cost of employing them. A major source of household income and a major cost to firms is compensation paid in exchange for labor services.

Additionally, households typically choose to spend less on consumption than they earn from their labor. This behavior is called **saving**, through which households can accumulate financial capital, the returns on which can produce other sources of household income, such as interest, dividends, and capital gains. Households may choose to lend their accumulated

savings (in exchange for interest) or invest it in ownership claims in firms (in hopes of receiving dividends and capital gains). Households make these savings choices when their anticipated future returns are judged to be more valuable today than the present consumption that households must sacrifice when they save.

Indeed, a major purpose of financial institutions and markets is to enable the transfer of these savings into capital investments. Firms use **capital markets** (markets for long-term financial capital—that is, markets for long-term claims on firms' assets and cash flows) to sell debt (in bond markets) or equity (in equity markets) in order to raise funds to invest in productive assets, such as plant and equipment. They make these investment choices when they judge that their investments will increase the value of the firm by more than the cost of acquiring those funds from households. Firms also use such financial intermediaries as banks and insurance companies to raise capital, typically debt funding that ultimately comes from the savings of households, which are usually net accumulators of financial capital.

Microeconomics, although primarily focused on goods and factor markets, can contribute to the understanding of all types of markets (e.g., markets for financial securities).

EXAMPLE 1-1 Types of Markets

1. Which of the following markets is *least* accurately described as a factor market? The market for:
   A. land.
   B. assembly-line workers.
   C. capital market securities.

2. Which of the following markets is *most* accurately defined as a product market? The market for:
   A. companies.
   B. unskilled labor.
   C. legal and lobbying services.

Solution to 1: C is correct.

Solution to 2: C is correct.

3. BASIC PRINCIPLES AND CONCEPTS

In this chapter, we explore a model of household behavior that yields the consumer demand curve. **Demand**, in economics, is the willingness and ability of consumers to purchase a given amount of a good or service at a given price. **Supply** is the willingness of sellers to offer a given quantity of a good or service for a given price. Later, study on the theory of the firm will yield the supply curve.

The demand and supply model is useful in explaining how price and quantity traded are determined and how external influences affect the values of those variables. Buyers' behavior is

captured in the demand function and its graphical equivalent, the demand curve. This curve shows both the highest price buyers are willing to pay for each quantity and the largest quantity buyers are willing and able to purchase at each price. Sellers' behavior is captured in the supply function and its graphical equivalent, the supply curve. This curve shows simultaneously the lowest price sellers are willing to accept for each quantity and the largest quantity sellers are willing to offer at each price.

If, at a given quantity, the highest price that buyers are willing to pay is equal to the lowest price that sellers are willing to accept, we say the market has reached its equilibrium quantity. Alternatively, when the quantity that buyers are willing and able to purchase at a given price is just equal to the quantity that sellers are willing to offer at that same price, we say the market has discovered the equilibrium price. So equilibrium price and quantity are achieved simultaneously, and as long as neither the supply curve nor the demand curve shifts, there is no tendency for either price or quantity to vary from its equilibrium value.

3.1. The Demand Function and the Demand Curve

We first analyze demand. The quantity consumers are willing to buy clearly depends on a number of different factors, called variables. Perhaps the most important of those variables is the item's own price. In general, economists believe that as the price of a good rises, buyers will choose to buy less of it, and as its price falls, they buy more. This is such a ubiquitous observation that it has come to be called the **law of demand**, although we shall see that it need not hold in all circumstances.

Although a good's own price is important in determining consumers' willingness to purchase it, other variables also have influence on that decision, such as consumers' incomes, their tastes and preferences, the prices of other goods that serve as substitutes or complements, and so on. Economists attempt to capture all of these influences in a relationship called the **demand function**. (In general, a function is a relationship that assigns a unique value to a dependent variable for any given set of values of a group of independent variables.) We represent such a demand function in Equation 1-1:

$$Q_x^d = f(P_x, I, P_y, \dots) \tag{1-1}$$

where Q_x^d represents the quantity demanded of some good X (such as per-household demand for gasoline in gallons per week), P_x is the price per unit of good X (such as \$ per gallon), I is consumers' income (as in \$1,000s per household annually), and P_y is the price of another good, Y. (There can be many other goods, not just one, and they can be complements or substitutes.) Equation 1-1 may be read, "Quantity demanded of good X depends on (is a function of) the price of good X, consumers' income, the price of good Y, and so on."

Often, economists use simple linear equations to approximate real-world demand and supply functions in relevant ranges. A hypothetical example of a specific demand function could be Equation 1-2, a linear equation for a small town's per-household gasoline consumption per week, where P_y might be the average price of an automobile in \$1,000s:

$$Q_x^d = 8.4 - 0.4P_x + 0.06I - 0.01P_y \tag{1-2}$$

The signs of the coefficients on gasoline price (negative) and consumer's income (positive) are intuitive, reflecting, respectively, an inverse and a positive relationship between those variables and quantity of gasoline consumed. The negative sign on average automobile price may indicate that if automobiles go up in price, fewer will be purchased and driven; hence less gasoline will be consumed. As will be discussed later, such a relationship would indicate that gasoline and automobiles have a negative cross-price elasticity of demand and are thus complements.

To continue our example, suppose that the price of gasoline (P_x) is \$3 per gallon, per-household income (I) is \$50,000, and the price of the average automobile (P_y) is \$20,000. Then this function would predict that the per-household weekly demand for gasoline would be 10 gallons: $8.4 - 0.4(3) + 0.06(50) - 0.01(20) = 8.4 - 1.2 + 3 - 0.2 = 10$, recalling that income and automobile prices are measured in thousands. Note that the sign on the own-price variable is negative; thus, as the price of gasoline rises, per-household weekly consumption would decrease by 0.4 gallons for every dollar increase in gas price. **Own-price** is used by economists to underscore that the reference is to the price of a good itself and not the price of some other good.

In our example, there are three independent variables in the demand function, and one dependent variable. If any one of the independent variables changes, so does the value of quantity demanded. It is often desirable to concentrate on the relationship between the dependent variable and just one of the independent variables at a time, which allows us to represent the relationship between those two variables in a two-dimensional graph (at specific levels of the variables held constant). To accomplish this goal, we can simply hold the other two independent variables constant at their respective levels and rewrite the equation. In economic writing, this "holding constant" of the values of all variables except those being discussed is traditionally referred to by the Latin phrase *ceteris paribus* (literally, "all other things being equal" in the sense of unchanged). In this chapter, we use the phrase "holding all other things constant" as a readily understood equivalent for *ceteris paribus*.

Suppose, for example, that we want to concentrate on the relationship between the quantity demanded of the good and its own price, P_x. Then we would hold constant the values of income and the price of good Y. In our example, those values are 50 and 20, respectively. So, by inserting the respective values, we would rewrite Equation 1-2 as:

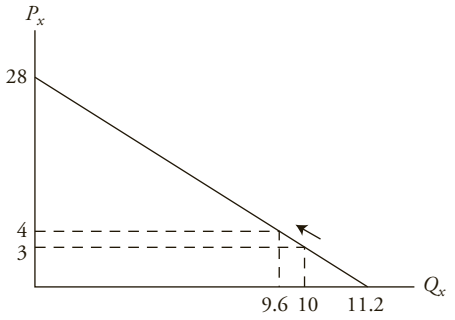
$$Q_x^d = 8.4 - 0.4P_x + 0.06(50) - 0.01(20) = 11.2 - 0.4P_x \tag{1-3}$$

Notice that income and the price of automobiles are not ignored; they are simply held constant, and they are collected in the new constant term, 11.2. Notice also that we can rearrange Equation 1-3, solving for P_x in terms of Q_x. This operation is called "inverting the demand function," and gives us Equation 1-4. (You should be able to perform this algebraic exercise to verify the result.)

$$P_x = 28 - 2.5Q_x \tag{1-4}$$

Equation 1-4, which gives the per-gallon price of gasoline as a function of gasoline consumed per week, is referred to as the **inverse demand function**. We need to restrict Q_x in Equation 1-4 to be less than or equal to 11.2 so price is not negative. Henceforward we assume that the reader can work out similar needed qualifications to the valid application of

EXHIBIT 1-1 Household Demand Curve for Gasoline



equations. The graph of the inverse demand function is called the **demand curve**, and is shown in Exhibit 1-1.[1]

This demand curve is drawn with price on the vertical axis and quantity on the horizontal axis. Depending on how we interpret it, the demand curve shows either the greatest quantity a household would buy at a given price or the highest price it would be willing to pay for a given quantity. In our example, at a price of \$3 per gallon households would each be willing to buy 10 gallons per week. Alternatively, the highest price they would be willing to pay for 10 gallons per week is \$3 per gallon. Both interpretations are valid, and we will be thinking in terms of both as we proceed. If the price were to rise by \$1, households would reduce the quantity they each bought by 0.4 units to 9.6 gallons. We say that the slope of the demand curve is $1/-0.4$, or -2.5. Slope is always measured as "rise over run," or the change in the vertical variable divided by the change in the horizontal variable. In this case, the slope of the demand curve is $\Delta P/\Delta Q$, where "Δ" stands for "the change in." The change in price was \$1, and it is associated with a change in quantity of negative 0.4.

3.2. Changes in Demand versus Movements along the Demand Curve

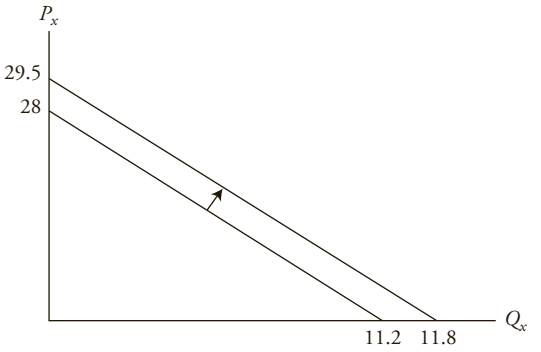
As we just saw, when own-price changes, quantity demanded changes. This change is called a movement along the demand curve or a change in quantity demanded, and it comes only from a change in own-price.

Recall that to draw the demand curve, though, we had to hold everything *except* quantity and own-price constant. What would happen if income were to change by some amount? Suppose that household income rose by \$10,000 per year to a value of 60. Then the value of Equation 1-3 would change to Equation 1-5:

$$Q_x^d = 8.4 - 0.4P_x + 0.06(60) - 0.01(20) = 11.8 - 0.4P_x \qquad (1\text{-}5)$$

[1]Following usual practice, here and in other exhibits we will show linear demand curves intersecting the quantity axis at a price of zero, which shows the intercept of the associated demand equation. Real-world demand functions may be nonlinear in some or all parts of their domain. Thus, linear demand functions in practical cases are viewed as approximations to the true demand function that are useful for a relevant range of values. The relevant range would typically not include a price of zero, and the prediction for demand at a price of zero should not be viewed as usable.

EXHIBIT 1-2 Household Demand Curve for Gasoline before and after Change in Income



and Equation 1-4 would become the new inverse demand function (Equation 1-6):

$$P_x = 29.5 - 2.5Q_x \tag{1-6}$$

Notice that the slope has remained constant, but the intercepts have both increased, resulting in an outward shift in the demand curve, as shown in Exhibit 1-2.

In general, the only thing that can cause a movement along the demand curve is a change in a good's own price. A change in the value of any *other* variable will shift the entire demand curve. The former is referred to as a *change in quantity demanded*, and the latter is referred to as a *change in demand*.

More importantly, the shift in demand was both a *vertical* shift upward and a *horizontal* shift to the right. That is to say, for any given quantity, the household is now willing to pay a higher price; and at any given price, the household is now willing to buy a greater quantity. Both interpretations of the shift in demand are valid.

EXAMPLE 1-2 Representing Consumer Buying Behavior with a Demand Function and Demand Curve

An individual consumer's monthly demand for downloadable e-books is given by the equation

$$Q_{eb}^d = 2 - 0.4P_{eb} + 0.0005I + 0.15P_{hb}$$

where Q_{eb}^d equals the number of e-books demanded each month, P_{eb} equals the price of e-books, I equals the household monthly income, and P_{hb} equals the price of

hardbound books, per unit. Notice that the sign on the price of hardbound books is positive, indicating that when hardbound books increase in price, more e-books are purchased; thus, according to this equation, the two types of books are substitutes. Assume that the price of each e-book is €10.68, household income is €2,300, and the price of each hardbound book is €21.40.

1. Determine the number of e-books demanded by this household each month.
2. Given the values for I and P_{hb}, determine the inverse demand function.
3. Determine the slope of the demand curve for e-books.
4. Calculate the vertical intercept (price-axis intercept) of the demand curve if income increases to €3,000 per month.

Solution to 1: Insert given values into the demand function and calculate quantity:

$$Q_{eb}^{d} = 2 - 0.4(10.68) + 0.0005(2{,}300) + 0.15(21.40) = 2.088$$

Hence, the household will demand e-books at the rate of 2.088 books per month. Note that this rate is a flow, so there is no contradiction in there being a noninteger quantity. In this case, the outcome means that the consumer buys 23 e-books during 11 months.

Solution to 2: We want to find the price–quantity relationship holding all other things constant, so first, insert values for I and P_{hb} into the demand function and collect the constant terms:

$$Q_{eb}^{d} = 2 - 0.4P_{eb} + 0.0005(2{,}300) + 0.15(21.40) = 6.36 - 0.4P_{eb}$$

Now solve for P_{eb} in terms of Q_{eb}: $P_{eb} = 15.90 - 2.5Q_{eb}$

Solution to 3: Note from the previous inverse demand function that when Q_{eb} rises by one unit, P_{eb} falls by €2.5. So the slope of the demand curve is -2.5, which is the coefficient on Q_{eb} in the inverse demand function. Note it is *not* the coefficient on P_{eb} in the demand function, which is -0.4. It is the inverse of that coefficient.

Solution to 4: In the demand function, change the value of I to 3,000 from 2,300 and collect constant terms:

$$Q_{eb}^{d} = 2 - 0.4P_{eb} + 0.0005(3{,}000) + 0.15(21.40) = 6.71 - 0.4P_{eb}$$

Now solve for P_{eb}: $P_{eb} = 16.78 - 2.5Q_{eb}$. The vertical intercept is 16.78. (Note that this increase in income has shifted the demand curve outward and upward but has not affected its slope, which is still -2.5.)

3.3. The Supply Function and the Supply Curve

The willingness and ability to sell a good or service is called supply. In general, producers are willing to sell their product for a price as long as that price is at least as high as the cost to produce an additional unit of the product. It follows that the willingness to supply, called the **supply function**, depends on the price at which the good can be sold as well as the cost of production for an additional unit of the good. The greater the difference between those two values, the greater is the willingness of producers to supply the good.

In subsequent chapters, we will explore the cost of production in greater detail. At this point, we need to understand only the basics of cost. At its simplest level, production of a good consists of transforming inputs, or factors of production (such as land, labor, capital, and materials), into finished goods and services. Economists refer to the rules that govern this transformation as the **technology of production**. Because producers have to purchase inputs in factor markets, the cost of production depends on both the technology and the price of those factors. Clearly, willingness to supply is dependent on not only the price of a producer's output, but additionally on the prices (i.e., costs) of the inputs necessary to produce it. For simplicity, we can assume that the only input in a production process is labor that must be purchased in the labor market. The price of an hour of labor is the wage rate, or W. Hence, we can say that (for any given level of technology) the willingness to supply a good depends on the price of that good and the wage rate. This concept is captured in Equation 1-7, which represents an individual seller's supply function:

$$Q_x^s = f(P_x, W, \dots) \tag{1-7}$$

where Q_x^s is the quantity supplied of some good X (such as gasoline), P_x is the price per unit of good X, and W is the wage rate of labor in, say, dollars per hour. It would be read, "The quantity supplied of good X depends on (is a function of) the price of X (its own price), the wage rate paid to labor, and so on."

Just as with the demand function, we can consider a simple hypothetical example of a seller's supply function. As mentioned earlier, economists often will simplify their analysis by using linear functions, although that is not to say that all demand and supply functions are necessarily linear. One hypothetical example of an individual seller's supply function for gasoline is given in Equation 1-8:

$$Q_x^s = -175 + 250P_x - 5W \tag{1-8}$$

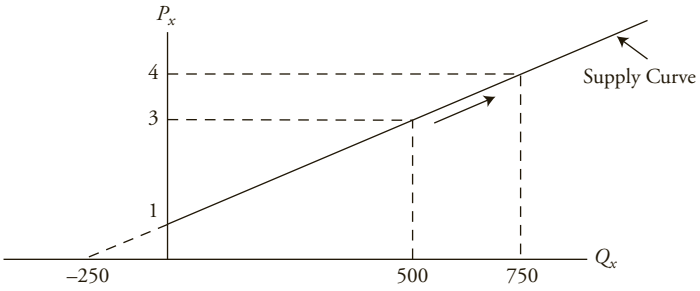
Notice that this supply function says that for every increase in price of \$1, this seller would be willing to supply an additional 250 units of the good. Additionally, for every \$1 increase in wage rate that it must pay its laborers, this seller would experience an increase in marginal cost and would be willing to supply five fewer units of the good.

We might be interested in the relationship between only two of these variables, price and quantity supplied. Just as we did in the case of the demand function, we use the assumption of ceteris paribus and hold everything except own-price and quantity constant. In our example, we accomplish this by setting W to some value, say, \$15. The result is Equation 1-9:

$$Q_x^s = -175 + 250P_x - 5(15) = -250 + 250P_x \tag{1-9}$$

in which only the two variables Q_x^s and P_x appear. Once again, we can solve this equation for P_x in terms of Q_x^s, which yields the *inverse supply function* in Equation 1-10:

EXHIBIT 1-3 Individual Seller's Supply Curve for Gasoline



$$P_x = 1 + 0.004Q_x \tag{1-10}$$

The graph of the inverse supply function is called the **supply curve**, and it shows simultaneously the highest quantity willingly supplied at each price and the lowest price willingly accepted for each quantity. For example, if the price of gasoline were \$3 per gallon, Equation 1-9 implies that this seller would be willing to sell 500 gallons per week. Alternatively, the lowest price the seller would accept and still be willing to sell 500 gallons per week would be \$3. Exhibit 1-3 represents our hypothetical example of an individual seller's supply curve of gasoline.

What does our supply function tell us will happen if the retail price of gasoline rises by \$1? We insert the new higher price of \$4 into Equation 1-8 and find that quantity supplied would rise to 750 gallons per week. The increase in price has enticed the seller to supply a greater quantity of gasoline per week than at the lower price.

3.4. Changes in Supply versus Movements along the Supply Curve

As we saw earlier, a change in the (own) price of a product causes a change in the quantity of that good willingly supplied. A rise in price typically results in a greater quantity supplied, and a lower price results in a lower quantity supplied. Hence, the supply curve has a positive slope, in contrast to the negative slope of a demand curve. This positive relationship is often referred to as the **law of supply**.

What happens when a variable other than own-price takes on different values? We could answer this question in our example by assuming a different value for wage rate, say \$20 instead of \$15. Recalling Equation 1-9, we would simply put in the higher wage rate and solve, yielding Equation 1-11.

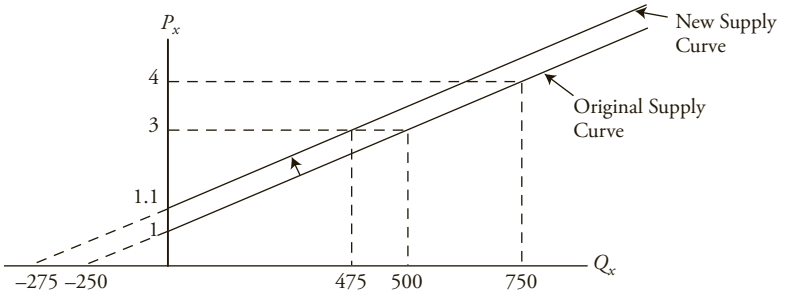
$$Q_x^s = -175 + 250P_x - 5(20) = -275 + 250P_x \tag{1-11}$$

This equation, too, can be solved for P_x, yielding the inverse supply function in Equation 1-12:

$$P_x = 1.1 + 0.004Q_x \tag{1-12}$$

Notice that the constant term has changed, but the slope has remained the same. The result is a shift in the entire supply curve, as illustrated in Exhibit 1-4.

EXHIBIT 1-4 Individual Seller's Supply Curve for Gasoline before and after Increase in Wage Rate



Notice that the supply curve has shifted both vertically upward and horizontally leftward as a result of the rise in the wage rate paid to labor. This change is referred to as a **change in supply**, as contrasted with a **change in quantity supplied** that would result only from a change in this product's own price. Now, at a price of 3, a lower quantity will be supplied: 475 instead of 500. Alternatively, in order to entice this seller to offer the same 500 gallons per week, the price would now have to be 3.1, up from 3 before the change. This increase in lowest acceptable price reflects the now higher marginal cost of production resulting from the increased input price that the firm now must pay for labor.

To summarize, a change in the price of a good itself will result in a movement along the supply curve and a change in quantity supplied. A change in any variable other than own-price will cause a shift in the supply curve, called a change in supply. This distinction is identical to the case of demand curves.

EXAMPLE 1-3 Representing Seller Behavior with a Supply Function and Supply Curve

An individual seller's monthly supply of downloadable e-books is given by the equation

$$Q^s_{eb} = -64.5 + 37.5P_{eb} - 7.5W$$

where Q^s_{eb} is number of e-books supplied each month, P_{eb} is price of e-books in euros, and W is the hourly wage rate in euros paid by e-book sellers to workers. Assume that the price of e-books is €10.68 and the hourly wage is €10.

1. Determine the number of e-books supplied each month.
2. Determine the inverse supply function for an individual seller.
3. Determine the slope of the supply curve for e-books.
4. Determine the new vertical intercept of the individual e-book supply curve if the hourly wage were to rise to €15 from €10.

Solution to 1: Insert given values into the supply function and calculate the number of e-books:

$$Q^s_{eb} = -64.5 + 37.5(10.68) - 7.5(10) = 261$$

Hence, each seller would be willing to supply e-books at the rate of 261 per month.

Solution to 2: Holding all other things constant, the wage rate is constant at €10, so we have:

$$Q^s_{eb} = -64.5 + 37.5P_{eb} - 7.5(10) = -139.5 + 37.5P_{eb}$$

We now solve this for P_{eb}:

$$P_{eb} = 3.72 + 0.0267Q_{eb}$$

Solution to 3: Note that when Q_{eb} rises by one unit, P_{eb} rises by 0.0267 euros, so the slope of the supply curve is 0.0267, which is the coefficient on Q_{eb} in the inverse supply function. Note that it is *not* 37.5.

Solution to 4: In the supply function, increase the value of W to €15 from €10:

$$Q^s_{eb} = -64.5 + 37.5P_{eb} - 7.5(15) = -177 + 37.5P_{eb}$$

and invert by solving for P_{eb}:

$$P_{eb} = 4.72 + 0.267Q_{eb}$$

The vertical intercept is now 4.72. Thus, an increase in the wage rate shifts the supply curve upward and to the left. This change is known as a decrease in supply because at each price the seller would be willing now to supply fewer e-books than before the increase in labor cost.

3.5. Aggregating the Demand and Supply Functions

We have explored the basic concept of demand and supply at the individual household and the individual supplier level. However, markets consist of collections of demanders and suppliers, so we need to understand the process of combining these individual agents' behavior to arrive at market demand and supply functions.

The process could not be more straightforward: simply add all the buyers together and add all the sellers together. Suppose there are 1,000 identical gasoline buyers in our hypothetical example, and they represent the total market. At, say, a price of $3 per gallon, we find that one household would be willing to purchase 10 gallons per week (when income and price of automobiles are held constant at $50,000 and $20,000, respectively). So, 1,000 identical buyers would be willing to purchase 10,000 gallons collectively. It follows that to aggregate 1,000 buyers' demand functions, simply multiply each buyer's quantity demanded by 1,000, as shown in Equation 1-13:

$$Q_x^d = 1{,}000(8.4 - 0.4P_x + 0.06I - 0.01P_y) = 8{,}400 - 400P_x + 60I - 10P_y \quad (1\text{-}13)$$

where Q_x^d represents the market quantity demanded. Note that if we hold I and P_y at their same respective values of 50 and 20 as before, we can collapse the constant terms and write the following Equation 1-14:

$$Q_x^d = 11{,}200 - 400P_x \quad (1\text{-}14)$$

Equation 1-14 is just Equation 1-3 (an individual household's demand function) multiplied by 1,000 households (Q_x^d represents thousands of gallons per week). Again, we can solve for P_x to obtain the market inverse demand function:

$$P_x = 28 - 0.0025Q_x \quad (1\text{-}15)$$

The market demand curve is simply the graph of the market inverse demand function, as shown in Exhibit 1-5.

It is important to note that the aggregation process sums all individual buyers' *quantities*, not the *prices* they are willing to pay; that is, we multiplied the demand function, *not* the inverse demand function, by the number of households. Accordingly, the market demand curve has the exact same price intercept as each individual household's demand curve. If, at a price of $28, a single household would choose to buy zero, then it follows that 1,000 identical households would choose, in aggregate, to buy zero as well. However, if each household chooses to buy 10 at a price of $3, then 1,000 identical households would choose to buy 10,000, as shown in Exhibit 1-5. Hence, we say that all individual demand curves *horizontally* (quantities), not *vertically* (prices), are added to arrive at the market demand curve.

Now that we understand the aggregation of demanders, the aggregation of suppliers is simple: We do exactly the same thing. Suppose, for example, that there are 20 identical sellers with the supply function given by Equation 1-8. To arrive at the market supply function, we simply multiply by 20 to obtain Equation 1-16:

EXHIBIT 1-5 Aggregate Weekly Market Demand for Gasoline as the Quantity Summation of All Households' Demand Curves

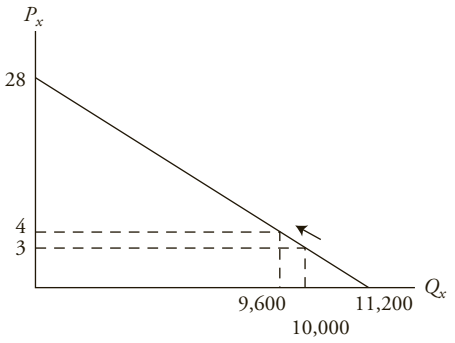
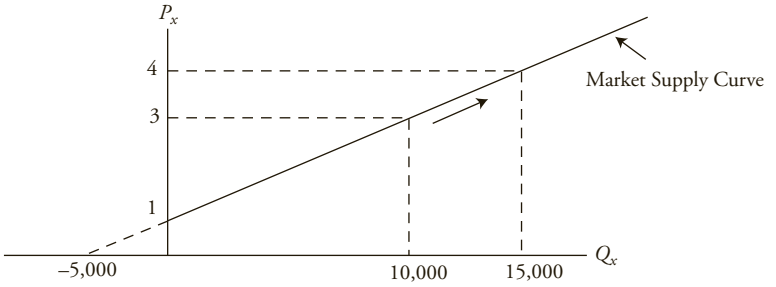


EXHIBIT 1-6 Aggregate Market Supply as the Quantity Summation of Individual Sellers' Supply Curves



$$Q_x^s = 20(-175 + 250P_x - 5W) = -3{,}500 + 5{,}000P_x - 100W \tag{1-16}$$

And, if we once again assume W equals \$15, we can collapse the constant terms, yielding Equation 1-17:

$$Q_x^s = 20[-175 + 250P_x - 5(15)] = -5{,}000 + 5{,}000P_x \tag{1-17}$$

which can be inverted to yield the market inverse supply function (Equation 1-18):

$$P_x = 1 + 0.0002Q_x \tag{1-18}$$

Graphing the market inverse supply function yields the market supply curve in Exhibit 1-6.

We saw from the individual seller's supply curve in Exhibit 1-3 that at a price of \$3, an individual seller would willingly offer 500 gallons of gasoline. It follows, as shown in Exhibit 1-6, that a group of 20 sellers would offer 10,000 gallons per week. Accordingly, at each price, the market quantity supplied is just 20 times as great as the quantity supplied by each seller. We see, as in the case of demand curves, that the market supply curve is simply the horizontal summation of all individual sellers' supply curves.

EXAMPLE 1-4 Aggregating Demand Functions

An individual consumer's monthly demand for downloadable e-books is given by the equation

$$Q_{eb}^d = 2 - 0.4P_{eb} + 0.0005I + 0.15P_{hb}$$

where Q_{eb}^d equals the number of e-books demanded each month, P_{eb} is the price of e-books in euros, I equals the household monthly income, and P_{hb} equals the price

of hardbound books per unit. Assume that household income is €2,300 and the price of hardbound books is €21.40. The market consists of 1,000 identical consumers with this demand function.

1. Determine the market aggregate demand function.
2. Determine the inverse market demand function.
3. Determine the slope of the market demand curve.

Solution to 1: Aggregating over the total number of consumers means summing up their demand functions (in the quantity direction). In this case, there are 1,000 consumers with identical individual demand functions, so multiply the entire function by 1,000:

$$\begin{aligned} Q_{eb} &= 1{,}000(2 - 0.4P_{eb} + 0.0005I + 0.15P_{hb}) \\ &= 2{,}000 - 400P_{eb} + 0.5I + 150P_{hb} \end{aligned}$$

Solution to 2: Holding I constant at a value of €2,300 and P_{hb} constant at a value of €21.40, we find

$$Q_{eb} = 2{,}000 - 400P_{eb} + 0.5(2{,}300) + 150(21.40) = 6{,}360 - 400P_{eb}$$

Now solve for $P_{eb} = 15.90 - 0.0025Q_{eb}$

Solution to 3: The slope of the market demand curve is the coefficient on Q_{eb} in the inverse demand function, which is -0.0025.

EXAMPLE 1-5 Aggregating Supply Functions

An individual seller's monthly supply of downloadable e-books is given by the equation

$$Q^s_{eb} = -64.5 + 37.5P_{eb} - 7.5W$$

where Q^s_{eb} is number of e-books supplied, P_{eb} is the price of e-books in euros, and W is the wage rate in euros paid by e-book sellers to laborers. Assume that the price of e-books is €10.68 and wage is €10. The supply side of the market consists of a total of eight identical sellers in this competitive market.

1. Determine the market aggregate supply function.
2. Determine the inverse market supply function.
3. Determine the slope of the aggregate market supply curve.

Solution to 1: Aggregating supply functions means summing up the quantity supplied by all sellers. In this case, there are eight identical sellers, so multiply the individual seller's supply function by eight:

$$Q^s_{eb} = 8(-64.5 + 37.5P_{eb} - 7.5W) = -516 + 300P_{eb} - 60W$$

Solution to 2: Holding W constant at a value of €10, insert that value into the aggregate supply function and then solve for P_{eb} to find the inverse supply function:

$$Q_{eb} = -1{,}116 + 300P_{eb}$$

Inverting, $P_{eb} = 3.72 + 0.0033Q_{eb}$

Solution to 3: The slope of the supply curve is the coefficient on Q_{eb} in the inverse supply function, which is 0.0033.

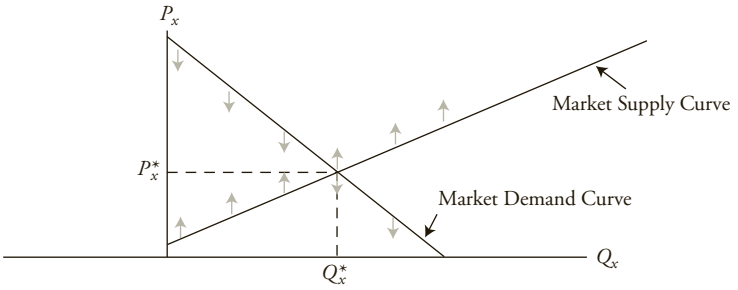
3.6. Market Equilibrium

An important concept in the market model is **market equilibrium**, defined as the condition in which the quantity willingly offered for sale by sellers at a given price is just equal to the quantity willingly demanded by buyers at that same price. When that condition is met, we say that the market has discovered its equilibrium price. An alternative and equivalent condition of equilibrium occurs at that quantity at which the highest price a buyer is willing to pay is just equal to the lowest price a seller is willing to accept for that same quantity.

As we have discovered in the earlier sections, the demand curve shows (for given values of income, other prices, etc.) an infinite number of combinations of prices and quantities that satisfy the demand function. Similarly, the supply curve shows (for given values of input prices, etc.) an infinite number of combinations of prices and quantities that satisfy the supply function. Equilibrium occurs at the unique combination of price and quantity that simultaneously satisfies *both* the market demand function and the market supply function. Graphically, it is the intersection of the demand and supply curves as shown in Exhibit 1-7.

In Exhibit 1-7, the shaded arrows indicate, respectively, that buyers will be willing to pay any price at or below the demand curve (indicated by ↓), and sellers are willing to accept any price at or above the supply curve (indicated by ↑).

EXHIBIT 1-7 Market Equilibrium Price and Quantity as the Intersection of Demand and Supply



Notice that for quantities less than Q_x^*, the highest price that buyers are willing to pay exceeds the lowest price that sellers are willing to accept, as indicated by the shaded arrows. But for all quantities above Q_x^*, the lowest price willingly accepted by sellers is greater than the highest price willingly offered by buyers. Clearly, trades will not be made beyond Q_x^*.

Algebraically, we can find the equilibrium price by setting the demand function equal to the supply function and solving for price. Recall that in our hypothetical example of a local gasoline market, the demand function was given by $Q_x^d = f(P_x, I, P_y)$, and the supply function was given by $Q_x^s = f(P_x, W)$. Those expressions are called **behavioral equations** because they model the behavior of, respectively, buyers and sellers. Variables *other* than own-price and quantity are determined outside of the demand and supply model of this particular market. Because of that, they are called **exogenous variables**. Price and quantity, however, are determined within the model for this particular market and are called **endogenous variables**. In our simple example, there are three exogenous variables (I, P_y, and W) and three endogenous variables: P_x, Q_x^d, and Q_x^s. Hence, we have a system of two equations and three unknowns. We need another equation to solve this system. That equation is called the **equilibrium condition**, and it is simply $Q_x^d = Q_x^s$.

Continuing with our hypothetical examples, we could assume that income equals \$50 (thousands, per year), the price of automobiles equals \$20 (thousands, per automobile), and the hourly wage equals \$15. In this case, our equilibrium condition can be represented in Equation 1-19 by setting Equation 1-14 equal to Equation 1-17:

$$11{,}200 - 400P_x = -5{,}000 + 5{,}000P_x \tag{1-19}$$

and solving for equilibrium, $P_x = 3$.

Equivalently, we could have equated the inverse demand function to the inverse supply function (Equations 1-15 and 1-18, respectively), as shown in Equation 1-20:

$$28 - 0.0025Q_x = 1 + 0.0002Q_x \tag{1-20}$$

and solved for equilibrium, $Q_x = 10{,}000$. That is to say, for the given values of I and W, the unique combination of price and quantity of gasoline that results in equilibrium is (3, 10,000).

Note that our system of equations requires explicit values for the exogenous variables to find a unique equilibrium combination of price and quantity. Conceptually, the values of the exogenous variables are being determined in other markets, such as the markets for labor, automobiles, and so on, whereas the price and quantity of gasoline are being determined in the gasoline market. When we concentrate on one market, taking values of exogenous variables as given, we are engaging in what is called **partial equilibrium analysis**. In many cases, we can gain sufficient insight into a market of interest without addressing feedback effects to and from all the other markets that are tangentially involved with this one. At other times, however, we need explicitly to take account of all the feedback mechanisms that are going on in all markets simultaneously. When we do that, we are engaging in what is called **general equilibrium analysis**. For example, in our hypothetical model of the local gasoline market, we recognize that the price of automobiles, a complementary product, has an impact on the demand for gasoline. If the price of automobiles were to rise, people would tend to buy fewer automobiles and probably buy less gasoline. Additionally, though, the price of gasoline probably has an impact on the demand for automobiles, which, in turn, can feed back to the gasoline market. Because we are positing a very local gasoline market, it is probably safe to ignore all the feedback effects, but if we are modeling the national markets for gasoline and automobiles, a general equilibrium model might be warranted.

EXAMPLE 1-6 Finding Equilibrium by Equating Demand and Supply

In the local market for e-books, the aggregate demand is given by the equation

$$Q^d_{eb} = 2{,}000 - 400P_{eb} + 0.5I + 150P_{hb}$$

and the aggregate supply is given by the equation

$$Q^s_{eb} = -516 + 300P_{eb} - 60W$$

where Q_{eb} is quantity of e-books, P_{eb} is the price of an e-book, I is household income, W is wage rate paid to e-book laborers, and P_{hb} is the price of a hardbound book. Assume I is €2,300, W is €10, and P_{hb} is €21.40. Determine the equilibrium price and quantity of e-books in this local market.

Solution: Market equilibrium occurs when quantity demanded is equal to quantity supplied, so set $Q^d_{eb} = Q^s_{eb}$ after inserting the given values for the exogenous variables:

$$2{,}000 - 400P_{eb} + 0.5(2{,}300) + 150(21.4) = -516 + 300P_{eb} - 60(10)$$
$$6{,}360 - 400P_{eb} = -1{,}116 + 300P_{eb}$$

which implies that $P_{eb} = €10.68$, and $Q_{eb} = 2{,}088$.

3.7. The Market Mechanism: Iterating toward Equilibrium—or Not

It is one thing to define equilibrium as we have done, but we should also understand the mechanism for reaching equilibrium. That mechanism is what takes place when the market is *not* in equilibrium. Consider our hypothetical example. We found that the equilibrium price was 3, but what would happen if, by some chance, price was actually equal to 4? To find out, we need to see how much buyers would demand at that price and how much sellers would offer to sell by inserting 4 into the demand function and into the supply function.

In the case of quantity demanded, we find that (Equation 1-21):

$$Q^d_x = 11{,}200 - 400(4) = 9{,}600 \tag{1-21}$$

and in the case of quantity supplied (Equation 1-22),

$$Q^s_x = -5{,}000 + 5{,}000(4) = 15{,}000 \tag{1-22}$$

Clearly, the quantity supplied is greater than the quantity demanded, resulting in a condition called **excess supply**, as illustrated in Exhibit 1-8. In our example, there are 5,400 more units of this good offered for sale at a price of 4 than are demanded at that price.

EXHIBIT 1-8 Excess Supply as a Consequence of Price above Equilibrium Price

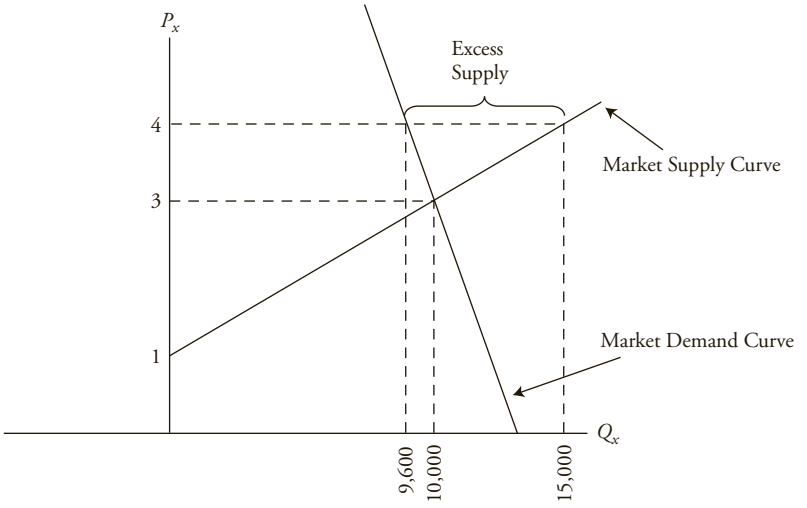
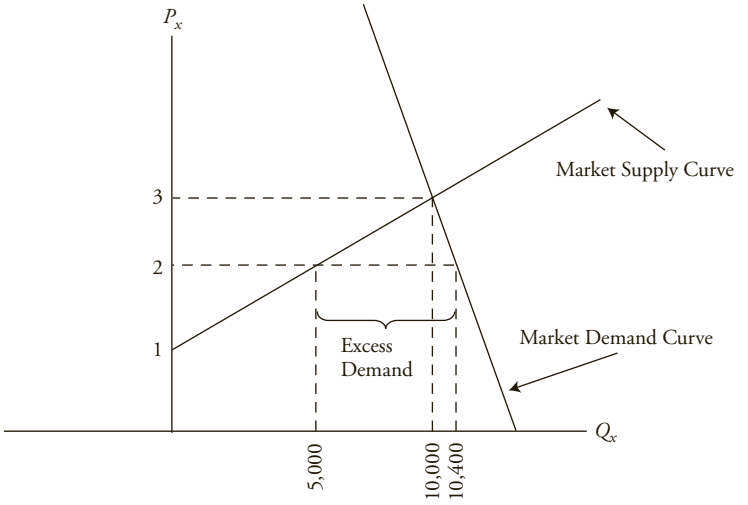


EXHIBIT 1-9 Excess Demand as a Consequence of Price below Equilibrium Price



Alternatively, if the market was presented with a price that was too low, say 2, then by inserting the price of 2 into Equations 1-21 and 1-22, we find that buyers are willing to purchase 5,400 *more* units than sellers are willing to offer. This result is shown in Exhibit 1-9.

To reach equilibrium, price must adjust until there is neither an excess supply nor an excess demand. That adjustment is called the **market mechanism**, and it is characterized in the following way: In the case of excess supply, price will fall; in the case of excess demand, price will rise; and in the case of neither excess supply nor excess demand, price will not change.

EXAMPLE 1-7 Identifying Excess Demand or Excess Supply at a Nonequilibrium Price

In the local market for e-books, the aggregate demand is given by the equation

$$Q^d_{eb} = 6{,}360 - 400P_{eb}$$

and the aggregate supply by the equation

$$Q^s_{eb} = -1{,}116 + 300P_{eb}$$

1. Determine the amount of excess demand or supply if price is €12.
2. Determine the amount of excess demand or supply if price is €8.

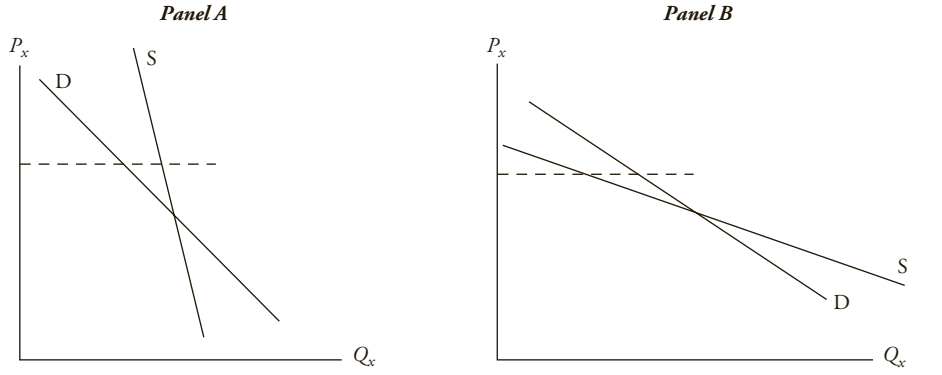
Solution to 1: Insert the presumed price of €12 into the demand function to find $Q^d_{eb} = 6{,}360 - 400(12) = 1{,}560$. Insert a price of €12 into the supply function to find $Q^s_{eb} = -1{,}116 - 300(12) = 2{,}484$. Because quantity supplied is greater than quantity demanded at the €12 price, there is an excess supply equal to $2{,}484 - 1{,}560 = 924$.

Solution to 2: Insert the presumed price of €8 into the demand function to find $Q^d_{eb} = 6{,}360 - 400(8) = 3{,}160$. Insert a price of €8 into the supply function to find $Q^s_{eb} = -1{,}116 + 300(8) = 1{,}284$. Because quantity demanded is greater than quantity supplied at the €8 price, there is an excess demand equal to $3{,}160 - 1{,}284 = 1{,}876$.

It might be helpful to consider the following process in our hypothetical market. Suppose that some neutral agent or referee were to display a price for everyone in the market to observe. Then, given that posted price, we would ask each potential buyer to write down on a slip of paper a quantity that he or she would be willing and able to purchase at that price. At the same time, each potential seller would write down a quantity that he or she would be willing to sell at that price. Those pieces of paper would be submitted to the referee, who would then calculate the total quantity demanded and the total quantity supplied at that price. If the two sums are identical, the slips of paper would essentially become contracts that would be executed, and the session would be concluded by buyers and sellers actually trading at that price. If there was an excess supply, however, the referee's job would be to discard the earlier slips of paper and display a price lower than before. Alternatively, if there was an excess demand at the original posted price, the referee would discard the slips of paper and post a higher price. This process would continue until the market reached an equilibrium price at which the quantity willingly offered for sale would just equal the quantity willingly purchased. In this way, the market could tend to move toward equilibrium.[2]

[2]The process described is known among economists as Walrasian *tâtonnement*, after the French economist Léon Walras (1834–1910). *Tâtonnement* means, roughly, "searching," referring to the mechanism for establishing the equilibrium price.

EXHIBIT 1-10 Stability of Equilibria: I



Note: If supply intersects demand from above, equilibrium is dynamically stable.

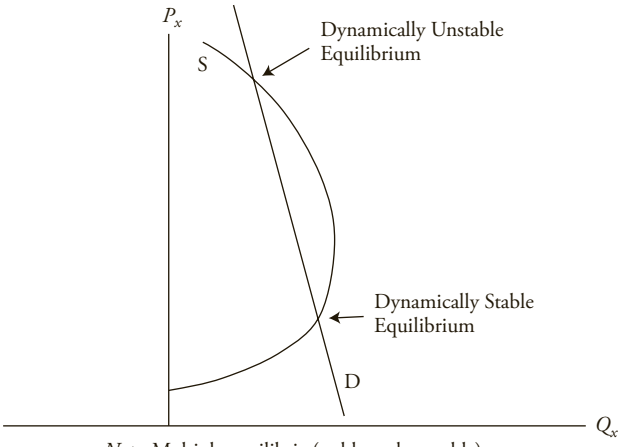
Note: If supply intersects demand from below, equilibrium is dynamically unstable.

It is not really necessary for a market to have such a referee for it to operate *as if* it had one. Experimental economists have simulated markets in which subjects (usually college students) are given an order either to purchase or to sell some amount of a commodity for a price either no higher (in the case of buyers) or no lower (in the case of sellers) than a set dollar limit. Those limits are distributed among market participants and represent a positively sloped supply curve and a negatively sloped demand curve. The goal for buyers is to buy at a price as far below their limit as possible, and the goal for sellers to sell at a price as far above their minimum as possible. The subjects are then allowed to interact in a simulated trading pit by calling out willingness to buy or sell. When two participants come to an agreement on a price, that trade is then reported to a recorder, who displays the terms of the deal. Traders are then allowed to observe current prices as they continue to search for a buyer or a seller. It has consistently been shown in experiments that this mechanism of open outcry buying and selling (historically, one of the oldest mechanisms used in trading securities) soon converges to the theoretical equilibrium price and quantity inherent in the underlying demand and supply curves used to set the respective sellers' and buyers' limit prices.

In our hypothetical example of the gasoline market, the supply curve is positively sloped, and the demand curve is negatively sloped. In that case, the market mechanism would tend to reach an equilibrium and return to that equilibrium whenever price was accidentally bumped away from it. We refer to such an equilibrium as being **stable** because whenever price is disturbed away from equilibrium, it tends to converge back to that equilibrium.[3] It is possible, however, for this market mechanism to result in an unstable equilibrium. Suppose that not only does the demand curve have a negative slope but also the supply curve has a negatively sloped segment. For example, at some level of wages, a wage increase might cause workers to supply fewer hours of work if satisfaction (utility) gained from an extra hour of leisure is greater than the satisfaction obtained from an extra hour of work. Then two possibilities could result, as shown in Panels A and B of Exhibit 1-10.

[3]In the same sense, equilibrium may sometimes also be referred to as being *dynamically stable*. Similarly, *unstable* or *dynamically unstable* may be used in the sense introduced later.

EXHIBIT 1-11 Stability of Equilibria: II



Note: Multiple equilibria (stable and unstable) can result from nonlinear supply curves.

Notice that in Panel A both demand (D) and supply (S) are negatively sloped, but S is steeper and intersects D from above. In this case, if price is above equilibrium, there will be excess supply and the market mechanism will adjust price downward toward equilibrium. In Panel B, D is steeper, which results in S intersecting D from below. In this case, at a price above equilibrium there will be excess demand, and the market mechanism will dictate that price should *rise*, thus leading away from equilibrium. This equilibrium would be considered **unstable**. If price were accidentally displayed above the equilibrium price, the mechanism would not cause price to converge to that equilibrium, but instead to soar above it because there would be excess demand at that price. In contrast, if price were accidentally displayed below equilibrium, the mechanism would force price even further below equilibrium because there would be excess supply.

If supply were nonlinear, there could be multiple equilibria, as shown in Exhibit 1-11.

Note that there are two combinations of price and quantity that would equate quantity supplied and demanded, hence two equilibria. The lower-priced equilibrium is stable, with a positively sloped supply curve and a negatively sloped demand curve. However, the higher-priced equilibrium is unstable because at a price above that equilibrium price there would be excess demand, thus driving price even higher. At a price below that equilibrium there would be excess supply, thus driving price even lower toward the lower-priced equilibrium, which is a stable equilibrium.

Observation suggests that most markets are characterized by stable equilibria. Prices do not often shoot off to infinity or plunge toward zero. However, occasionally we do observe price bubbles occurring in real estate, securities, and other markets. It appears that prices can behave in ways that are not ultimately sustainable in the long run. They may shoot up for a time, but ultimately, if they do not reflect actual valuations, the bubble can burst, resulting in a correction to a new equilibrium.

As a simple approach to understanding bubbles, consider a case in which buyers and sellers base their expectations of future prices on the rate of change of current prices: if price rises, they take that as a sign that price will rise even further. Under these circumstances, if

buyers see an increase in price today, they might actually shift the demand curve to the right, desiring to buy more at each price today because they expect to have to pay more in the future. Alternately, if sellers see an increase in today's price as evidence that price will be even higher in the future, they are reluctant to sell today as they hold out for higher prices tomorrow, and that would shift the supply curve to the left. With a rightward shift in demand and a leftward shift in supply, buyers' and sellers' expectations about price are confirmed and the process begins again. This scenario could result in a bubble that would inflate until someone decides that such high prices can no longer be sustained. The bubble bursts and price plunges.

3.8. Auctions as a Way to Find Equilibrium Price

Sometimes markets really do use auctions to arrive at equilibrium price. Auctions can be categorized into two types depending on whether the value of the item being sold is the same for each bidder or is unique to each bidder. The first case is called a **common value auction** in which there is some actual common value that will ultimately be revealed after the auction is settled. Prior to the auction's settlement, however, bidders must estimate that true value. An example of a common value auction would be bidding on a jar containing many coins. Each bidder could estimate the value; but until someone buys the jar and actually counts the coins, no one knows with certainty the true value. In the second case, called a **private value auction**, each buyer places a subjective value on the item, and in general their values differ. An example might be an auction for a unique piece of art that buyers are hoping to purchase for their own personal enjoyment, not primarily as an investment to be sold later.

Auctions also differ according to the mechanism used to arrive at a price and to determine the ultimate buyer. These mechanisms include the ascending price (or English) auction, the first price sealed-bid auction, the second price sealed-bid (or Vickery) auction, and the descending price (or Dutch) auction.

Perhaps the most familiar auction mechanism is the **ascending price auction** in which an auctioneer is selling a single item in a face-to-face arena where potential buyers openly reveal their willingness to buy the good at prices that are called out by an auctioneer. The auctioneer begins at a low price and easily elicits nods from buyers. He or she then raises the price incrementally. In a common value auction, buyers can sometimes learn something about the true value of the item being auctioned from observing other bidders. Ultimately bidders with different maximum amounts they are willing to pay for the item, called **reservation prices**, begin to drop out of the bidding as price rises above their respective reservation prices.[4] Finally, only one bidder is left (who has outbid the bidder with the second-highest valuation) and the item is sold to that bidder for that bid price.

Sometimes sellers offer a common value item, such as an oil or timber lease, in a **sealed-bid auction**. In this case, bids are elicited from potential buyers, but there is no ability to observe bids by other buyers until the auction has ended. In the **first price sealed-bid auction**, the envelopes containing bids are opened simultaneously and the item is sold to the highest bidder for the actual highest bid price. Consider an oil lease being auctioned by the government. The highest bidder will pay the bid price but does not know with certainty the profitability of the asset being bid on. The profits that are ultimately realized will be learned only after a successful bidder buys and exploits the asset. Bidders each have some expected value they place on the oil lease, and those values can vary among bidders. Typically, some overly optimistic bidders will

[4]The term *reservation price* is also used to refer to the minimum price the seller of the auctioned item is willing to accept.

value the asset higher than its ultimate realizable value, and they might submit bids above that true value. Because the highest bidder wins the auction and must pay the full bid price, the highest bidder may fall prey to the **winner's curse** of having bid more than the ultimate value of the asset. The winner in this case will lose money because of having paid more than the value of the asset being auctioned. In recognition of the possibility of being overly optimistic, bidders might bid very conservatively below their expectation of the true value. If all bidders react in this way, the seller might end up with a low sale price.

If the item being auctioned is a private value item, then there is no danger of the winner's curse (no one would bid more than their own true valuation). But bidders try to guess the reservation prices of other bidders, so the most successful winning bidder would bid a price just above the reservation price of the second-highest bidder. This bid will be below the true reservation price of the highest bidder, resulting in a "bargain" for the highest bidder. To induce each bidder to reveal their true reservation price, sellers can use the **second price sealed-bid** mechanism (also known as a Vickery auction). In this mechanism, the bids are submitted in sealed envelopes and opened simultaneously. The winning buyer is the one who submitted the highest bid, but the price paid is not equal to the winner's own bid. The winner pays a price equal to the second-highest bid. The optimal strategy for bidders in such an auction is to bid their actual reservation prices, so the second price sealed-bid auction induces buyers to reveal their true valuation of the item. It is also true that if the bidding increments are small, the second price sealed-bid auction will yield the same ultimate price as the ascending price auction.

Yet another type of auction is called a **descending price auction** or **Dutch auction** in which the auctioneer begins at a very high price—a price so high that no bidder is believed to be willing to pay it.[5] The auctioneer then lowers the called price in increments until there is a willing buyer of the item being sold. If there are many bidders, each with a different reservation price and a unit demand, then each has a perfectly vertical demand curve at one unit and a height equal to his or her reservation price. For example, suppose the highest reservation price is equal to $100. That person would be willing to buy one unit of the good at a price no higher than $100. Suppose each subsequent bidder also has a unit demand and a reservation price that falls, respectively, in increments of $1. The market demand curve would be a negatively sloped step function; that is, it would look like a stair step, with the width of each step being one unit and the height of each step being $1 lower than the preceding step. For example, at a price equal to $90, 11 people would be willing to buy one unit of the good. If the price were to fall to $89, then the quantity demanded would be 12, and so on.

In the Dutch auction, the auctioneer would begin with a price above $100 and then lower it by increments until the bidder with the highest reservation price would purchase the unit. Again, the supply curve for this single-unit auction would be vertical at one unit, although there might be a seller reserve price that would form the lower bound on the supply curve at that reserve price.

A traditional Dutch auction as just described could be conducted in a single-unit or multiple-unit format. With a multiple-unit format, the price quoted by the auctioneer would be the per-unit price and a winning bidder could take fewer units than all the units for sale. If the winning bidder took fewer than all units for sale, the auctioneer would then lower the price until all units for sale were sold; thus transactions could occur at multiple prices. Modified Dutch auctions (frequently also called simply Dutch auctions in practice) are

[5]The historical use of this auction type for flower auctions in the Netherlands explains the name.

commonly used in securities markets; the modifications often involve establishing a single price for all purchasers. As implemented in share repurchases, the company stipulates a range of acceptable prices at which the company would be willing to repurchase shares from existing shareholders. The auction process is structured to uncover the minimum price at which the company can buy back the desired number of shares, with the company paying that price to all qualifying bids. For example, if the share price is €25 per share, the company might offer to repurchase three million shares in a range of €26 to €28 per share. Each shareholder would then indicate the number of shares and the lowest price at which he or she would be willing to sell. The company would then begin to qualify bids beginning with those shareholders who submitted bids at €26 and continue to qualify bids at higher prices until three million shares had been qualified. In our example, that price might be €27. Shareholders who bid between €26 and €27, inclusive, would then be paid €27 per share for their shares.

Another Dutch auction variation, also involving a single price and called a **single price auction**, is used in selling U.S. Treasury securities.[6] The single price Treasury bill auction operates as follows: The Treasury announces that it will auction 26-week T-bills with an offering amount of, say, $90 billion with both competitive and noncompetitive bidding. Noncompetitive bidders state the total face value they are willing to purchase at the ultimate price (yield) that clears the market (i.e., sells all of the securities offered), whatever that turns out to be. Competitive bidders each submit a total face value amount and the price at which they are willing to purchase those T-bills. The Treasury then ranks those bids in ascending order of yield (i.e., descending order of price) and finds the yield at which the total $90 billion offering amount would be sold. If the offering amount is just equal to the total face value bidders are willing to purchase at that yield, then all the T-bills are sold for that single yield. If there is excess demand at that yield, then bidders would each receive a proportionately smaller total than they offered.

As an example, suppose the following table shows the prices and the offers from competitive bidders for a variety of prices, as well as the total offers from noncompetitive bidders, assumed to be $15 billion:

Discount Rate Bid (%)	Bid Price per $100	Competitive Bids ($ billions)	Cumulative Competitive Bids ($ billions)	Noncompetitive Bids ($ billions)	Total Cumulative Bids ($ billions)
0.1731	99.91250	10	10	15	25
0.1741	99.91200	15	25	15	40
0.1751	99.91150	20	45	15	60
0.1760	99.91100	12	57	15	72
0.1770	99.91050	10	67	15	82
0.1780	99.91000	5	72	15	87
0.1790	99.90950	10	82	15	97

[6]Historically, the U.S. Treasury has also used multiple price auctions, and in the euro area multiple price auctions are widely used. See www.dsta.nl/english/Subjects/Auction_methods for more information.

At yields below 0.1790 percent (prices above 99.90950), there is still excess supply. But at that yield, more bills are demanded than the $90 billion face value of the total offer amount. The clearing yield would be 0.1790 percent (a price of 99.9095 per $100 of face value), and all sales would be made at that single yield. All the noncompetitive bidders would have their orders filled at the clearing price, as well as all bidders who bid above that price. The competitive bidders who offered a price of 99.9095 would have 30 percent of their orders filled at that price because it would take only 30 percent of the $10 billion ($90 billion − $87 billion offered = $3 billion, or 30 percent of $10 billion) demanded at that price to complete the $90 billion offer amount. That is, by filling 30 percent of the competitive bids at a price of 99.9095, the cumulative competitive bids would sum to $75 billion. This amount plus the $15 billion noncompetitive bids adds up to $90 billion.

EXAMPLE 1-8 Auctioning Treasury Bills with a Single Price Auction

The U.S. Treasury offers to sell $115 billion of 52-week T-bills and requests competitive and noncompetitive bids. Noncompetitive bids total $10 billion, and competitive bidders in descending order of offer price are as given in the table:

Discount Rate Bid (%)	Bid Price per $100	Competitive Bids ($ billions)	Cumulative Competitive Bids ($ billions)	Noncompetitive Bids ($ billions)	Total Cumulative Bids ($ billions)
0.1575	99.8425	12			
0.1580	99.8420	20			
0.1585	99.8415	36			
0.1590	99.8410	29			
0.1595	99.8405	5			
0.1600	99.8400	15			
0.1605	99.8395	10			

1. Determine the winning price if a single price Dutch auction is used to sell these T-bills.
2. For those bidders at the winning price, what percentage of their order would be filled?

Solution to 1: Enter the noncompetitive quantity of $10 billion into the table. Then find the cumulative competitive bids and the total cumulative bids in the respective columns:

Bid Price per $100	Competitive Bids ($ billions)	Cumulative Competitive Bids ($ billions)	Noncompetitive Bids ($ billions)	Total Cumulative Bids ($ billlions)
99.8425	12	12	10	22
99.8420	20	32	10	42
99.8415	36	68	10	78
99.8410	29	97	10	107
99.8405	5	102	10	112
99.8400	15	117	10	127
99.8395	10	127	10	137

Note that at a bid price of 99.8400 there would be excess demand of $12 billion (i.e., the difference between $127 billion bid and $115 billion offered), but at the higher price of 99.8405 there would be excess supply. So the winning bid would be at a price of 99.8400.

Solution to 2: At a price of 99.8400, there would be $15 billion more demanded than at 99.8405 ($127 billion minus $112 billion), and at 99.8405 there would be excess supply equal to $3 billion. So the bidders at the winning bid would have only 3/15, or 20 percent, of their orders filled.

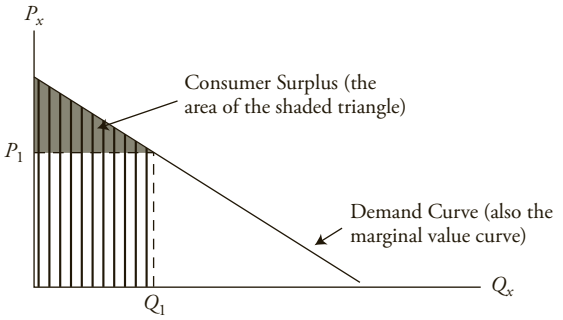
3.9. Consumer Surplus—Value minus Expenditure

To this point, we have discussed the fundamentals of demand and supply curves and explained a simple model of how a market can be expected to arrive at an equilibrium combination of price and quantity. While it is certainly necessary for the analyst to understand the basic workings of the market model, it is also crucial to have a sense of why we might care whether the market tends toward equilibrium. This question moves us into the normative, or evaluative, consideration of whether market equilibrium is desirable in any social sense. In other words, is there some reasonable measure we can apply to the outcome of a competitive market that enables us to say whether that outcome is socially desirable? Economists have developed two related concepts called consumer surplus and producer surplus to address that question. We will begin with consumer surplus, which is a measure of how much net benefit buyers enjoy from the ability to participate in a particular market.

To get an intuitive feel for this concept, consider the last thing you purchased. Maybe it was a cup of coffee, a new pair of shoes, or a new car. Whatever it was, think of how much you actually paid for it. Now contrast that price with the maximum amount you would have been *willing to pay* for it instead of going without it altogether. If those two numbers are different, we say you received some consumer surplus from your purchase. You received a bargain because you were willing to pay more than you had to pay.

Earlier we referred to the law of demand, which says that as price falls, consumers are willing to buy more of the good. This observation translates into a negatively sloped demand

EXHIBIT 1-12 Consumer Surplus



Note: Consumer surplus is the area beneath the demand curve and above the price paid.

curve. Alternatively, we could say that the highest price that consumers are willing to pay for an additional unit declines as they consume more and more of it. In this way, we can interpret their *willingness to pay* as a measure of how much they *value* each additional unit of the good. This point is very important: To purchase a unit of some good, consumers must give up something else they value. So the price they are willing to pay for an additional unit of a good is a measure of how much they value that unit, in terms of the other goods they must sacrifice to consume it.

If demand curves are negatively sloped, it must be because the value of each additional unit of the good falls the more of it they consume. We will explore this concept further later, but for now it is enough to recognize that the demand curve can thus be considered a **marginal value curve** because it shows the highest price consumers are willing to pay for each *additional* unit. In effect, the demand curve is the willingness of consumers to pay for each additional unit.

This interpretation of the demand curve allows us to measure the total value of consuming any given quantity of a good: It is the sum of all the marginal values of each unit consumed, up to and including the last unit. Graphically, this measure translates into the area under the consumer's demand curve, up to and including the last unit consumed, as shown in Exhibit 1-12, in which the consumer is choosing to buy Q_1 units of the good at a price of P_1. The **marginal value** of the Q_1th unit is clearly P_1, because that is the highest price the consumer is willing to pay for that unit. Importantly, however, the marginal value of each unit *up to* the Q_1th unit is greater than P_1.[7]

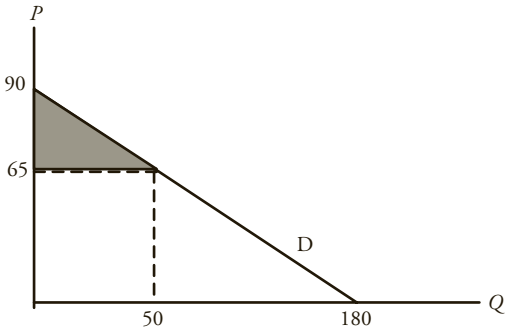
Because the consumer would have been willing to pay more for each of those units than she actually paid (P_1), then we can say she received more value than the cost to her of buying them. This concept is referred to as **consumer surplus**, and it is defined as the difference between the value that the consumer places on those units and the amount of money that was required to pay for them. The total value of Q_1 is thus the area of the vertically lined trapezoid in Exhibit 1-12. The **total expenditure** is only the area of the rectangle with height P_1 and base Q_1. The total consumer surplus received from buying Q_1 units at a level price of P_1 per unit is the *difference* between the area under the demand curve, on the one hand, and the area of the rectangle, $P_1 \times Q_1$, on the other hand. That area is shown as the lightly shaded triangle.

[7]This assumes that all units of the good are sold at the same price, P_1. Because the demand curve is negatively sloped, all units up to the Q_1th have marginal values greater than that price.

EXAMPLE 1-9 Calculating Consumer Surplus

A market demand function is given by the equation $Q^d = 180 - 2P$. Determine the value of consumer surplus if price is equal to 65.

Solution: First, insert 65 into the demand function to find the quantity demanded at that price: $Q^d = 180 - 2(65) = 50$. Then, to make drawing the demand curve easier, invert the demand function by solving it for P in terms of Q: $P = 90 - 0.5Q$. Note that the price intercept is 90, and the quantity intercept is 180. Draw the demand curve:



Find the area of the triangle above the price and below the demand curve, up to quantity 50. Area of a triangle is given as ½ Base × Height = (½)(50)(25) = 625.

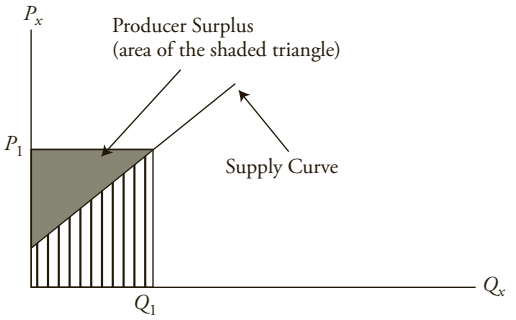
3.10. Producer Surplus—Revenue minus Variable Cost

In this section, we discuss a concept analogous to consumer surplus called **producer surplus**. It is the difference between the total revenue sellers receive from selling a given amount of a good, on the one hand, and the total variable cost of producing that amount, on the other hand. **Variable costs** are those costs that change when the level of output changes. Total revenue is simply the total quantity sold multiplied by the price per unit.

The total variable cost (variable cost per unit times units produced) is measured by the area beneath the supply curve, and it is a little more complicated to understand. Recall that the supply curve represents the lowest price that sellers would be willing to accept for each additional unit of a good. In general, that amount is the cost of producing that next unit, called **marginal cost**. Clearly, a seller would never intend to sell a unit of a good for a price *lower* than its marginal cost, because the seller would lose money on that unit. Alternatively, a producer should be more than willing to sell that unit for a price that is *higher* than its marginal cost, because it would contribute something toward fixed cost and profit, and obviously the higher the price the better for the seller. Hence, we can interpret the marginal cost curve as the lowest price sellers would accept for each quantity, which basically means that the marginal cost curve is the supply curve of any competitive seller. The market supply curve is simply the aggregation of all sellers' individual supply curves, as we showed in section 3.5.

Marginal cost curves are likely to have positive slopes. (It is the logical result of the law of diminishing marginal product, which will be discussed in a later chapter.) In Exhibit 1-13, we see

EXHIBIT 1-13 Producer Surplus



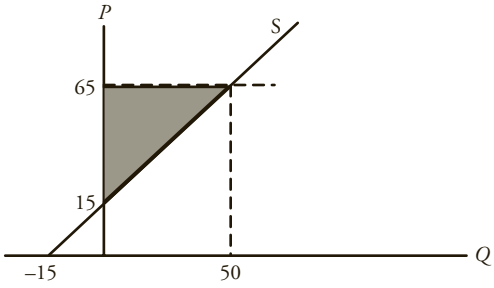
Note: Producer surplus is the area beneath the price and above the supply curve.

such a supply curve. Because its height is the marginal cost of each additional unit, the total variable cost of Q_1 units is measured as the area beneath the supply curve, up to and including that Q_1th unit, or the area of the vertically lined trapezoid. But each unit is being sold at the same price P_1, so total revenue to sellers is the rectangle whose height is P_1 and base is total quantity Q_1. Because sellers would have been willing to accept the amount of money represented by the trapezoid but they actually received the larger area of the rectangle, we say they received producer surplus equal to the area of the shaded triangle. So sellers also got a bargain because they received a higher price than they would have been willing to accept for each unit.

EXAMPLE 1-10 Calculating Producer Surplus

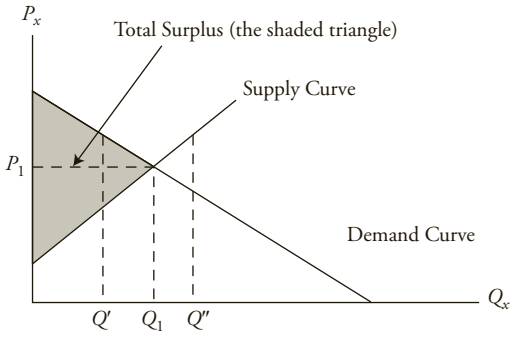
A market supply function is given by the equation $Q^s = -15 + P$. Determine the value of producer surplus if price were equal to 65.

Solution: First, insert 65 into the supply function to find the quantity supplied at that price: $Q^s = -15 + (65) = 50$. Then, to make drawing the supply curve easier, invert the supply function by solving for P in terms of Q: $P = 15 + Q$. Note that the price intercept is 15, and the quantity intercept is -15. Draw the supply curve:



Find the area of the triangle below the price and above the supply curve, up to a quantity of 50. Area = ½ Base × Height = (½)(50)(50) = 1,250.

EXHIBIT 1-14 Total Surplus as the Area beneath the Demand Curve and above the Supply Curve



3.11. Total Surplus—Total Value minus Total Variable Cost

In the previous sections, we have seen that consumers and producers both receive a bargain when they are allowed to engage in a mutually beneficial, voluntary exchange with one another. For every unit up to the equilibrium unit traded, buyers would have been willing to pay more than they ended up actually having to pay. Additionally, for every one of those units, sellers would have been willing to sell it for less than they actually received. The total value to buyers was greater than the total variable cost to sellers. The difference between those two values is called **total surplus**, and it is made up of the sum of consumer surplus and producer surplus. Note that the way the total surplus is divided between consumers and producers depends on the steepness of the demand and supply curves. If the supply curve is steeper than the demand curve, more of the surplus is being captured by producers. If the demand curve is steeper, consumers capture more of the surplus.

In a fundamental sense, total surplus is a measure of society's gain from the voluntary exchange of goods and services. Whenever total surplus increases, society gains. An important result of market equilibrium is that total surplus is maximized at the equilibrium price and quantity. Exhibit 1-14 combines the supply curve and the demand curve to show market equilibrium and total surplus, represented as the area of the shaded triangle. The area of that triangle is the difference between the trapezoid of total value to society's buyers and the trapezoid of total resource cost to society's sellers. If price measures dollars (or euros) per unit, and quantity measures units per month, then the measure of total surplus is dollars (euros) per month. It is the bargain that buyers and sellers together experience when they voluntarily trade the good in a market. If the market ceased to exist, that would be the monetary value of the loss to society.

3.12. Markets Maximize Society's Total Surplus

Recall that the market demand curve can be considered the willingness of consumers to pay for each additional unit of a good. Hence, it is society's marginal value curve for that good. Additionally, the market supply curve represents the marginal cost to society to produce each additional unit of that good, assuming no positive or negative externalities. An **externality** is a case in which production costs or the consumption benefits of a good or service spill over onto

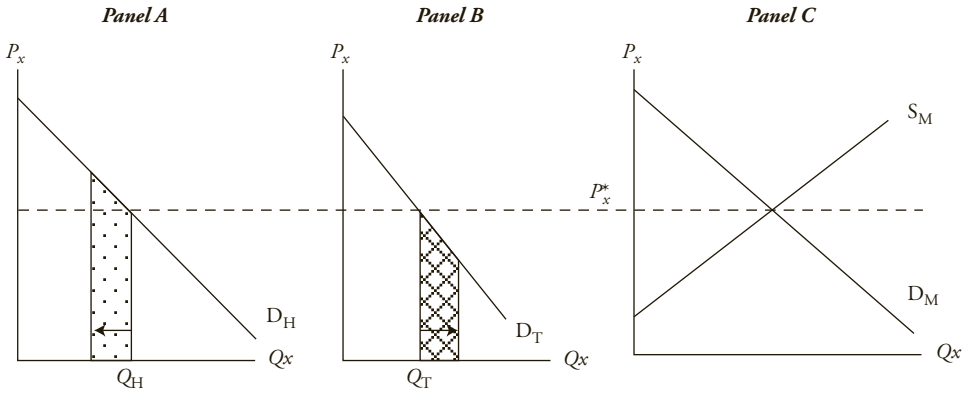
those who are not producing or consuming the good or service; a spillover cost (e.g., pollution) is called a **negative externality**, a spillover benefit (e.g., literacy programs) is called a **positive externality**.

At equilibrium, where demand and supply curves intersect, the highest price that someone is willing to pay is just equal to the lowest price that a seller is willing to accept, which is the marginal cost of that unit of the good. In Exhibit 1-14, that equilibrium quantity is Q_1. Now, suppose that some influence on the market caused less than Q_1 units to be traded, say only Q' units. Note that the marginal value of the Q'th unit exceeds society's marginal cost to produce it. In a fundamental sense, we could say that society *should* produce and consume it, as well as the next, and the next, all the way up to Q_1. Or suppose that some influence caused more than Q_1 to be produced, say Q'' units. Then what can we say? For all those units beyond Q_1 and up to Q'', society incurred greater cost than the value it received from consuming them. We could say that society *should not* have produced and consumed those additional units. Total surplus was reduced by those additional units because they cost more in the form of resources than the value they provided for society when they were consumed.

There is reason to believe that markets usually trend toward equilibrium and that the condition of equilibrium itself is also optimal in a welfare sense. To delve a little more deeply, consider two consumers, Helen Smith and Tom Warren, who have access to a market for some good, perhaps gasoline or shoes or any other consumption good. We could depict their situations using their individual demand curves juxtaposed on an exhibit of the overall market equilibrium, as in Exhibit 1-15 where Smith's and Warren's individual demands for a particular good are depicted along with the market demand and supply of that same good. (The horizontal axes are scaled differently because the market quantity is so much greater than either consumer's quantity, but the price axes are identical.)

At the market price of P_x^*, Smith chooses to purchase Q_H, and Warren chooses to purchase Q_T because at that price, the marginal value for each of the two consumers is just equal to the price they have to pay per unit. Now, suppose someone removed one unit of the good from Smith and presented it to Warren. In Panel A of Exhibit 1-15, the loss of value experienced by Smith is depicted by the dotted trapezoid, and in Panel B of Exhibit 1-15, the

EXHIBIT 1-15 How Total Surplus Can Be Reduced by Rearranging Quantity



Note: Beginning at a competitive market equilibrium, when one unit is taken from Smith and presented to Warren, total surplus is reduced.

gain in value experienced by Warren is depicted by the crosshatched trapezoid. Note that the increase in Warren's value is necessarily less than the loss in Smith's. Recall that consumer surplus is value minus expenditure. Total consumer surplus is reduced when individuals consume quantities that do not yield equal marginal value to each one. Conversely, when all consumers face the identical price, they will purchase quantities that equate their marginal values across all consumers. Importantly, that behavior maximizes total consumer surplus.

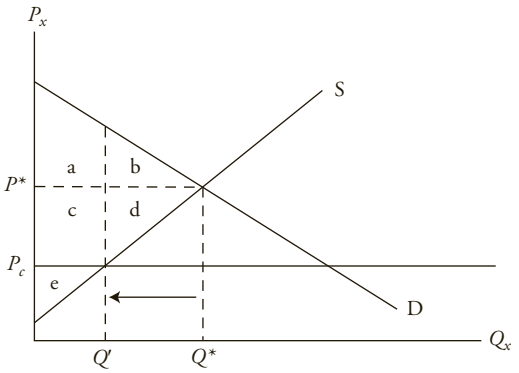
A precisely analogous argument can be made to show that when all producers produce quantities such that their marginal costs are equated across all firms, total producer surplus is maximized. The result of this analysis is that when all consumers face the same market equilibrium price and are allowed to buy all they desire at that price, and when all firms face that same price and are allowed to sell as much as they want at that price, the total of consumer and producer surplus (total surplus) is maximized from that market. This result is the beauty of free markets: They maximize society's net benefit from production and consumption of goods and services.

3.13. Market Interference: The Negative Impact on Total Surplus

Sometimes, lawmakers determine that the market price is too high for consumers to pay, so they use their power to impose a ceiling on price below the market equilibrium price. Some examples of ceilings include rent controls (limits on increases in the rent paid for apartments), limits on the prices of medicines, and laws against price gouging after a hurricane (i.e., charging opportunistically high prices for goods such as bottled water or plywood). Certainly, price limits benefit anyone who had been paying the old higher price and can still buy all they want but at the lower ceiling price. However, the story is more complicated than that. Exhibit 1-16 shows a market in which a ceiling price, P_c, has been imposed below equilibrium. Let's examine the full impact of such a law.

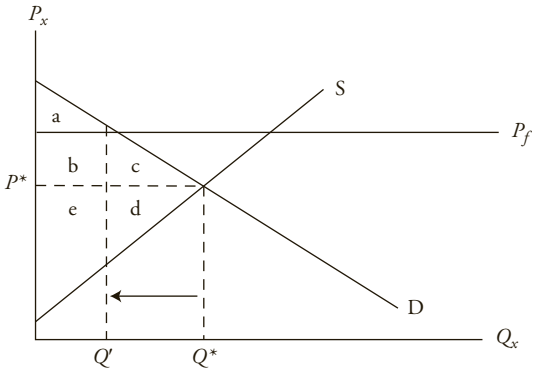
Prior to imposition of the ceiling price, equilibrium occurs at (P^*, Q^*), and total surplus equals the area given by a + b + c + d + e. It consists of consumer surplus given by a + b, and producer surplus given by c + d + e. When the ceiling is imposed, two things happen: Buyers

EXHIBIT 1-16 A Price Ceiling



Note: A price ceiling transfers surplus equal to area c from sellers to buyers, but it destroys surplus equal to area b + d, called a deadweight loss.

EXHIBIT 1-17 A Price Floor



Note: A price floor transfers surplus equal to area b from buyers to sellers, but it destroys surplus equal to area c + d, called a deadweight loss.

would like to purchase more at the lower price, but sellers are willing now to sell less. Regardless of how much buyers would like to purchase, though, only Q' would be offered for sale. Clearly, the total quantity that actually gets traded has fallen, and this has some serious consequences. For one thing, any buyer who is still able to buy the Q' quantity has clearly been given a benefit. Buyers used to pay P^* and now pay only P_c per unit. Those buyers gain consumer surplus equal to rectangle c, which used to be part of seller surplus. Rectangle c is surplus that has been transferred from sellers to buyers, but it still exists as part of total surplus. Disturbingly, though, there is a loss of consumer surplus equal to triangle b and a loss of producer surplus equal to triangle d. Those measures of surplus simply no longer exist at the lower quantity. Clearly, surplus cannot be enjoyed on units that are neither produced nor consumed, so that loss of surplus is called a **deadweight loss** because it is surplus that is lost by one or the other group but not transferred to anyone. Thus, after the imposition of a price ceiling at P_c, consumer surplus is given by a + c, producer surplus by e, and the deadweight loss is b + d.[8]

Another example of price interference is a **price floor**, in which lawmakers make it illegal to buy or sell a good or service below a certain price, which is above equilibrium. Again, some sellers who are still able to sell at the now higher floor price benefit from the law, but that's not the whole story. Exhibit 1-17 shows such a floor price, imposed at P_f above free market equilibrium.

At free market equilibrium quantity Q^*, total surplus is equal to a + b + c + d + e, consisting of consumer surplus equal to area a + b + c, and producer surplus equal to area e + d. When the floor is imposed, sellers would like to sell more, but buyers would choose to purchase less. Regardless of how much producers want to sell, however, only Q' will be purchased at the new higher floor price. Those sellers who can still sell at the higher price benefit at the expense of the buyers: There is a transfer of surplus from buyers to sellers equal to rectangle b.

[8]Technically, the statement assumes that the limited sales are allocated to the consumers with the highest valuations. A detailed explanation, however, is outside the scope of this chapter.

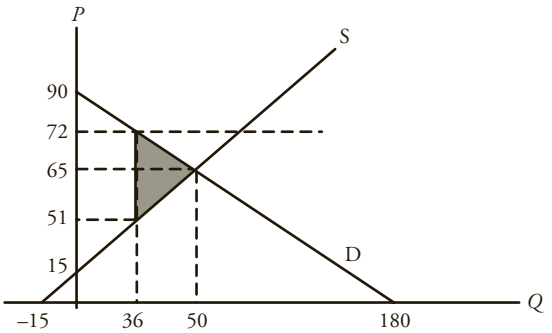
Regrettably, however, that's not all. Buyers also lose consumer surplus equal to triangle c, and sellers lose producer surplus equal to triangle d.[9] Once again, no one can benefit from units that are neither produced nor consumed, so there is a deadweight loss equal to triangle c plus triangle d. As a result of the floor, the buyer surplus is reduced to triangle a.

A good example of a price floor is the imposition of a legal minimum wage in the United States, the United Kingdom, and many other countries. Although controversy remains among some economists on the empirical effects of the minimum wage, most economists continue to believe that a minimum wage can reduce employment. Although some workers will benefit because they continue to work, now at the higher wage, others will be harmed because they will no longer be working at all.

EXAMPLE 1-11 Calculating the Amount of Deadweight Loss from a Price Floor

A market has demand function given by the equation $Q^d = 180 - 2P$, and supply function given by the equation $Q^s = -15 + P$. Calculate the amount of deadweight loss that would result from a price floor imposed at a level of 72.

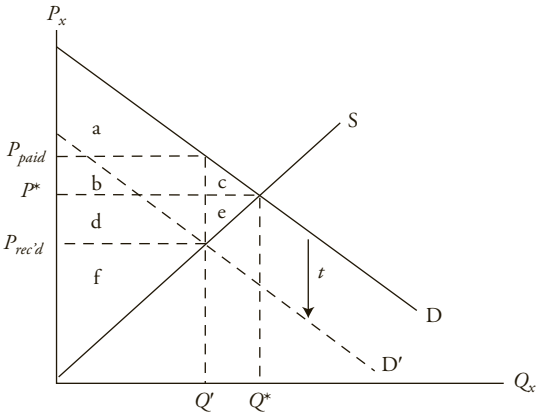
Solution: First, solve for equilibrium price of 65 and quantity 50. Then, invert the demand function to find $P = 90 - 0.5Q$, and the supply function to find $P = 15 + Q$. Use these functions to draw the demand and supply curves:



Insert the floor price of 72 into the demand function to find that only 36 would be demanded at that price. Insert 36 into the supply function to find the price of 51 that corresponds to a quantity of 36. Because the price floor would reduce quantity from its equilibrium value of 50 to the new value of 36, the deadweight loss would occur because those 14 units are not now being produced and consumed under the price floor. So deadweight loss equals the area of the shaded triangle: ½ Base × Height = (½)(72 − 51)(50 − 36) = 147.

[9]Technically, this statement assumes that sales are made by the lowest-cost producers. A discussion of the point is outside the scope of this chapter.

EXHIBIT 1-18 A Per-Unit Tax on Buyers

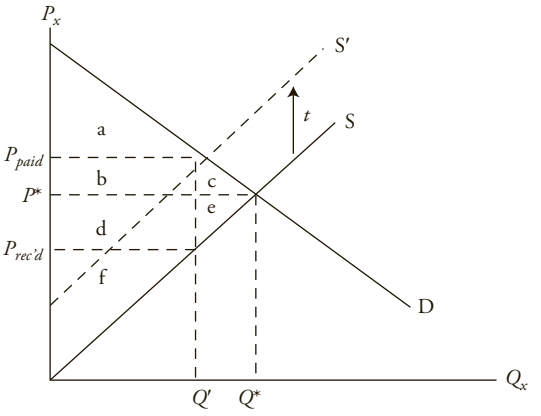


Note: A tax on buyers shifts the demand curve downward by *t*, imposing a burden on both buyers and sellers, shifting some of the surplus to government but leaving a deadweight loss equal to c plus e.

Still other policies can interfere with the ability of prices to allocate society's resources. Governments do have legitimate functions to perform in society, and they need to have revenue to finance them. So they often raise revenue by imposing taxes on various goods or activities. One such policy is a per-unit tax, such as an excise tax. By law, this tax could be imposed either on buyers or on sellers, but we will see that it really doesn't matter at all who legally must pay the tax; the result is the same: more deadweight loss. Exhibit 1-18 depicts such a tax imposed in this case on buyers. Here, the law simply says that whenever a buyer purchases a unit of some good, he or she must pay a tax of some amount *t* per unit. Recall that the demand curve is the highest price willingly paid for each quantity. Because buyers probably do not really care who receives the money, government or the seller, their gross willingness to pay is still the same. Because they must pay *t* dollars to the government, however, their net demand curve would shift vertically downward by *t* per unit. Exhibit 1-18 shows the result of such a shift.

Originally, the pretax equilibrium is where D and S intersect at (P^*, Q^*). Consumer surplus is given by triangle a plus rectangle b plus triangle c, and producer surplus consists of triangle f plus rectangle d plus triangle e. When the tax is imposed, the demand curve shifts vertically downward by the tax per unit, *t*. This shift results in a new equilibrium at the intersection of S and D′. That new equilibrium price is received by sellers ($P_{rec'd}$). However, buyers now must pay an additional *t* per unit to government, resulting in a total price paid (P_{paid}) that is higher than before. Sellers receive a lower price and buyers pay a higher price than pretax, so both suffer a burden as a result of this tax, even though it was legally imposed only on buyers. Buyers now have consumer surplus that has been reduced by rectangle b plus triangle c; thus, posttax consumer surplus is $(a + b + c) - (b + c) = a$. Sellers now have producer surplus that has been reduced by rectangle d plus triangle e; thus posttax producer surplus is $(f + d + e) - (d + e) = f$. Government receives tax revenue of *t* per unit multiplied

EXHIBIT 1-19 A Per-Unit Tax on Sellers



Note: A tax on sellers shifts the supply curve upward by *t*. Everything is exactly the same as in the case of imposing the tax on buyers.

by Q' units. Its total revenue is rectangle b plus rectangle d. Note that the total loss to buyers and sellers (b + c + d + e) is greater than the revenue transferred to government (b + d), so the tax resulted in a deadweight loss equal to triangle c plus triangle e as (b + c + d + e) − (b + d) = c + e.

How would things change if the tax had legally been imposed on sellers instead of buyers? To see the answer, note that the supply curve is the lowest price willingly accepted by sellers, which is their marginal cost. If they now must pay an additional *t* dollars per unit to government, their lowest acceptable price for each unit is now higher. We show this by shifting the supply curve vertically upward by *t* dollars per unit, as shown in Exhibit 1-19.

The new equilibrium occurs at the intersection of S′ and D, resulting in the new equilibrium price paid by buyers, P_{paid}. Sellers are paid this price but must remit *t* dollars per unit to the government, resulting in an after-tax price received ($P_{rec'd}$) that is lower than before the tax. In terms of overall result, absolutely nothing is different from the case in which buyers had the legal responsibility to pay the tax. Tax revenue to the government is the same, buyers' and sellers' reduction in surplus is identical to the previous case, and the deadweight loss is the same as well.

Notice that the share of the total burden of the tax need not be equal for buyers and sellers. In our example, sellers experienced a greater burden than buyers did, regardless of who had the legal responsibility to pay the tax. The relative burden from a tax falls disproportionately on the group (buyers or sellers) that has the steeper curve. In our example, the demand curve is flatter than the supply curve (just slightly so), so buyers bore proportionately less of the burden. Just the reverse would be true if the demand curve had been steeper than the supply curve.

All of the policies we have examined involve government interfering with free markets. Other examples include imposing tariffs on imported goods, setting quotas on imports, or banning the trade of goods. Additionally, governments often impose regulations on the production or consumption of goods to limit or correct the negative effects on third parties

that cannot be captured in free market prices. Even the most ardent of free market enthusiasts recognize the justification of some government intervention in the case of public goods, such as for national defense, or where prices do not reflect true marginal social value or cost, as in externalities such as pollution. Social considerations can trump pure economic efficiency, as in the case of child labor laws or human trafficking. What does come from the analysis of markets, however, is the recognition that when social marginal benefits are truly reflected in market demand curves and social marginal costs are truly reflected in supply curves, total surplus is maximized when markets are allowed to operate freely. Moreover, when society does choose to impose legal restrictions, market analysis of the kind we have just examined provides society with a means of at least assessing the deadweight losses that such policies extract from total surplus. In that way, policy makers can perform logical, rigorous cost-benefit assessments of their proposed policies to inform their decisions.

EXAMPLE 1-12 Calculating the Effects of a Per-Unit Tax on Sellers

A market has a demand function given by the equation $Q^d = 180 - 2P$, and a supply function given by the equation $Q^s = -15 + P$, where price is measured in euros per unit. A tax of €2 per unit is imposed on sellers in this market.

1. Calculate the effect on the price paid by buyers and the price received by sellers.
2. Demonstrate that the effect would be unchanged if the tax had been imposed on buyers instead of sellers.

Solution to 1: Determine the pretax equilibrium price and quantity by equating supply and demand: $180 - 2P = -15 + P$. Therefore $P^* = €65$ before tax. If the tax is imposed on sellers, the supply curve will shift upward by €2. So, to begin, we need to invert the supply function and the demand function: $P = 15 + Q^s$ and $P = 90 - 0.5Q^d$. Now, impose the tax on sellers by increasing the value of P by €2 at each quantity. This step simply means increasing the price intercept by €2. Because sellers must pay €2 tax per unit, the lowest price they are willing to accept for each quantity rises by that amount: $P' = 17 + Q^s$, where "P prime" indicates the new function after imposition of the tax. Because the tax was not imposed on buyers, the inverse demand function remains as it was. Solve for the new equilibrium price and quantity: $90 - 0.5Q = 17 + Q$, so new after-tax $Q = 48.667$. By inserting that quantity into the new inverse demand function, we find that $P_{paid} = €65.667$. This amount is paid by buyers to sellers, but because sellers are responsible for paying the €2 tax, they receive only €65.667 − €2 = €63.667 after tax. So we find that the tax on sellers has increased the price to buyers by €0.667 while reducing the price received by sellers by €1.33. Out of the €2 tax, buyers bear one-third of the burden and sellers bear two-thirds of the burden. This result is because the demand curve is half as steep as the supply curve. The group with the steeper, less elastic curve bears the greater burden of a tax, regardless of which group must legally pay the tax.

Solution to 2: Instead of adding €2 to the price intercept of the supply curve, we now subtract €2 from the price intercept of the demand curve. This step is because buyers' willingness to pay sellers has been reduced by the €2 they must pay in tax per unit. Buyers really don't care who receives their money; they are interested only in the greatest amount they are willing to pay for each quantity. So the new inverse demand function is $P'' = 88 - 0.5Q$. Using this new inverse demand, we now solve for equilibrium: $88 - 0.5Q = 15 + Q$. (Because buyers must pay the tax, we leave the old supply curve unchanged.) The new equilibrium quantity is therefore $Q = 48.667$, which is exactly as it was when sellers had the obligation to pay the tax. Inserting that number into the old supply function gives us the new equilibrium price of €63.667, which is what buyers must pay sellers. Recall, however, that now buyers must pay €2 in tax per unit, so the price buyers pay after tax is €63.667 + €2 = €65.667. So nothing changes when we impose the statutory obligation on buyers instead of sellers. They still share the ultimate burden of the tax in exactly the same proportion as when sellers had to send the €2 to the taxing authority.

We have seen that government interferences, such as price ceilings, price floors, and taxes, result in imbalances between demand and supply. In general, anything else that intervenes in the process of buyers and sellers finding the equilibrium price can cause imbalances as well. In the simple model of demand and supply, it is assumed that buyers and sellers can interact without cost. Often, however, there can be costs associated with finding a buyer's or a seller's counterpart. There could be a buyer who is willing to pay a price higher than some seller's lowest acceptable price, but if the two cannot find one another, there will be no transaction, resulting in a deadweight loss. The costs of matching buyers with sellers are generally referred to as **search costs**, and they arise because of frictions inherent in the matching process. When these costs are significant, an opportunity may arise for a third party to provide a valuable service by reducing those costs. This role is played by brokers. Brokers do not actually become owners of a good or service that is being bought or sold, but they serve the role of locating buyers for sellers or sellers for buyers. (Dealers, however, actually take possession of the item in anticipation of selling it to a future buyer.) To the extent that brokers serve to reduce search costs, they provide value in the transaction, and for that value they are able to charge a brokerage fee. Although the brokerage fee could certainly be viewed as a transaction cost, it is really a price charged for the service of reducing search costs. In effect, any impediment in the dissemination of information about buyers' and sellers' willingness to exchange goods can cause an imbalance in demand and supply. So anything that improves that information flow can add value. In that sense, advertising can add value to the extent that it informs potential buyers of the availability of goods and services.

4. DEMAND ELASTICITIES

The general model of demand and supply can be highly useful in understanding directional changes in prices and quantities that result from shifts in one or the other curve. At a deeper quantitative level, though, we often need to measure just *how* sensitive quantities demanded or supplied are to changes in the independent variables that affect them. Here is where the

concept of *elasticity of demand and supply* plays a crucial role in microeconomics. We will examine several elasticities of demand, but the crucial element is that fundamentally all elasticities are calculated the same way: they are ratios of percentage changes. Let us begin with the sensitivity of quantity demanded to changes in the own-price.

4.1. Own-Price Elasticity of Demand

Recall that when we introduced the concept of a demand function with Equation 1-1 earlier, we were simply theorizing that quantity demanded of some good, such as gasoline, is dependent on several other variables, one of which is the price of gasoline itself. We referred to the law of demand that simply states the inverse relationship between the quantity demanded and the price. Although that observation is useful, we might want to dig a little deeper and ask just how sensitive quantity demanded is to changes in the price of gasoline. Is the quantity demanded highly sensitive, so that a very small rise in price is associated with an enormous fall in quantity, or is the sensitivity only minimal? It might be helpful if we had a convenient measure of this sensitivity.

In Equation 1-3, we introduced a hypothetical household demand function for gasoline, assuming that the household's income and the price of another good (automobiles) were held constant. It supposedly described the purchasing behavior of a household regarding its demand for gasoline. That function was given by the simple linear expression $Q_x^d = 11.2 - 0.4P_x$. If we were to ask how sensitive quantity is to changes in price in that expression, one plausible answer would be simply to recognize that, according to that demand function, whenever price changes by one unit, quantity changes by 0.4 units in the opposite direction. That is to say, if price were to rise by \$1, quantity would fall by 0.4 gallons per week, so the coefficient on the price variable (-0.4) could be the measure of sensitivity we are seeking.

There is a fundamental drawback, however, associated with that measure. Notice that the -0.4 is measured in gallons of gasoline per dollar of price. It is crucially dependent on the *units* in which we measured Q and P. If we had measured the price of gasoline in cents per gallon instead of dollars per gallon, then the exact same household behavior would be described by the alternative equation $Q_x^d = 11.2 - 0.004P_x$. So, although we could choose the coefficient on price as our measure of sensitivity, we would always need to recall the units in which Q and P were measured when we wanted to describe the sensitivity of gasoline demand. That could be cumbersome.

Because of this drawback, economists prefer to use a gauge of sensitivity that does not depend on units of measure. That metric is called **elasticity**, and it is defined as the ratio of *percentage changes.* It is a general measure of how sensitive one variable is to any other variable. For example, if some variable y depends on some other variable x in the following function: $y = f(x)$, then the elasticity of y with respect to x is defined to be the percentage change in y divided by the percentage change in x, or $\%\Delta y/\%\Delta x$. In the case of **own-price elasticity of demand**, that measure is Equation 1-23:[10]

$$E_{P_x}^d = \frac{\%\Delta Q_x^d}{\%\Delta P_x} \tag{1-23}$$

[10]The reader will also encounter the Greek letter epsilon (ε) being used in the notation for elasticities.

Notice that this measure is independent of the units in which quantity and price are measured. If, for example, when price rises by 10 percent, quantity demanded falls by 8 percent, then elasticity of demand is simply -0.8. It does not matter whether we are measuring quantity in gallons per week or liters per day, and it does not matter whether we measure price in dollars per gallon or euros per liter; 10 percent is 10 percent, and 8 percent is 8 percent. So the ratio of the first to the second is still -0.8.

We can expand Equation 1-23 algebraically by noting that the percentage change in any variable x is simply the change in x (denoted "Δx") divided by the level of x. So, we can rewrite Equation 1-23, using a couple of simple steps, as Equation 1-24:

$$E^d_{P_x} = \frac{\%\Delta Q^d_x}{\%\Delta P_x} = \frac{\frac{\Delta Q^d_x}{Q^d_x}}{\frac{\Delta P_x}{P_x}} = \left(\frac{\Delta Q^d_x}{\Delta P_x}\right)\left(\frac{P_x}{Q^d_x}\right) \tag{1-24}$$

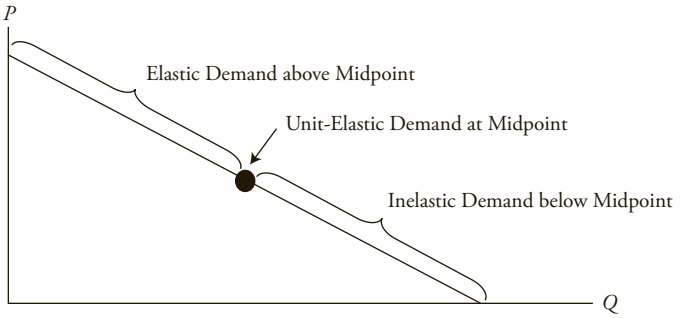
To get a better idea of price elasticity, it might be helpful to use our hypothetical market demand function: $Q^d_x = 11{,}200 - 400P_x$. For linear demand functions, the first term in the last line of Equation 1-24 is simply the slope coefficient on P_x in the demand function, or -400. (Technically, this term is the first derivative of Q^d_x with respect to P_x, dQ^d_x/dP_x, which is the slope coefficient for a linear demand function.) So, the elasticity of demand in this case is -400 multiplied by the ratio of price to quantity. Clearly in this case, we need to choose a price at which to calculate the elasticity coefficient. Let's choose the original equilibrium price of \$3. Now, we need to find the quantity associated with that particular price by inserting 3 into the demand function and finding $Q = 10{,}000$. The result of our calculation is that at a price of 3, the elasticity of our market demand function is $-400(3/10{,}000) = -0.12$. How do we interpret that value? It means, simply, that when price equals 3, a 1 percent rise in price would result in a fall in quantity demanded of only 0.12 percent. (You should try calculating price elasticity when price is equal to, say, \$4. Do you find that elasticity equals -0.167?)

In our particular example, when price is \$3 per gallon, demand is not very sensitive to changes in price, because a 1 percent rise in price would reduce quantity demanded by only 0.12 percent. Actually, that is not too different from empirical estimates of the actual demand elasticity for gasoline in the United States. When demand is not very sensitive to price, we say demand is **inelastic**. To be precise, when the *magnitude* (ignoring algebraic sign) of the own-price elasticity coefficient has a value less than 1, demand is defined to be inelastic. When that magnitude is greater than 1, demand is defined to be **elastic**. And when the elasticity coefficient is equal to negative 1, demand is said to be **unit elastic**, or **unitary elastic**. Note that if the law of demand holds, own-price elasticity of demand will always be negative, because a rise in price will be associated with a fall in quantity demanded, but it can be either elastic or inelastic. In our hypothetical example, suppose the price of gasoline was very high, say \$15 per gallon. In this case, the elasticity coefficient would be -1.154. Therefore, because the magnitude of the elasticity coefficient is greater than 1, we would say that demand is elastic at that price.[11]

By examining Equation 1-24, we should be able to see that for a linear demand curve the elasticity depends on where we calculate it. Note that the first term, $\Delta Q/\Delta P$, will remain constant along the entire demand curve because it is simply the inverse of the slope of the

[11]For evidence on price elasticities of demand for gasoline, see Espey (1996). The robust estimates were about -0.26 for short-run elasticity (less than one year) and -0.58 for more than a year.

EXHIBIT 1-20 The Elasticity of a Linear Demand Curve



Note: For all negatively sloped, linear demand curves, elasticity varies depending on where it is calculated.

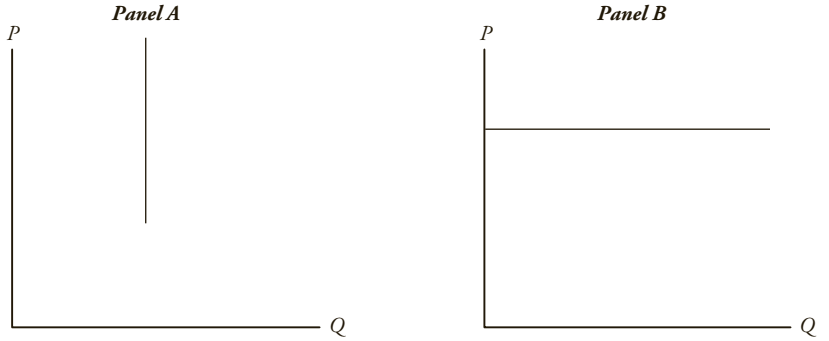
demand curve. But the second term, P/Q, clearly changes depending on where we look. At very low prices, P/Q is very small, so demand is inelastic. But at very high prices, Q is low and P is high, so the ratio P/Q is very high, and demand is elastic. Exhibit 1-20 illustrates a characteristic of all negatively sloped linear demand curves. Above the midpoint of the curve, demand is elastic; below the midpoint, demand is inelastic; and at the midpoint, demand is unit elastic.

Sometimes, we might not have the entire demand function or demand curve, but we might have just two observations on price and quantity. In this case, we do not know the slope of the demand curve at a given point because we really cannot say that it is even a linear function. For example, suppose we know that when price is 5, quantity demanded is 9,200, and when price is 6, quantity demanded is 8,800, but we do not know anything more about the demand function. Under these circumstances, economists use something called **arc elasticity**. Arc elasticity of demand is still defined as the percentage change in quantity demanded divided by the percentage change in price. However, because the choice of base for calculating percentage changes has an effect on the calculation, economists have chosen to use the *average* quantity and the *average* price as the base for calculating the percentage changes. (Suppose, for example, that you are making a wage of €10 when your boss says, "I'll increase your wage by 10 percent." You are then earning €11. But later that day, if your boss then reduces your wage by 10 percent, you are then earning €9.90. So, by receiving first a 10 percent raise and then a 10 percent cut in wage, you are worse off. The reason for this is that we typically use the original value as the base, or denominator, for calculating percentages.) In our example, then, the arc elasticity of demand would be:

$$E = \frac{\frac{\Delta Q}{Q_{avg}}}{\frac{\Delta P}{P_{avg}}} = \frac{\frac{-400}{9{,}000}}{\frac{1}{5.5}} = -0.244$$

There are two special cases in which linear demand curves have the same elasticity at all points: vertical demand curves and horizontal demand curves. Consider a vertical demand curve, as in Exhibit 1-21 Panel A, and a horizontal demand curve, as in Panel B. In the first case, the quantity demanded is the same regardless of price. Certainly, there could be no

EXHIBIT 1-21 The Extremes of Price Elasticity



Note: A vertical demand has zero elasticity and is called perfectly inelastic.

Note: A horizontal demand has infinite elasticity and is called perfectly elastic.

demand curve that is perfectly vertical at *all* possible prices, but over some range of prices it is not unreasonable that the same quantity would be purchased at a slightly higher price or a slightly lower price. Perhaps an individual's demand for, say, mustard might obey this description. Obviously, in that price range, quantity demanded is not at all sensitive to price and we would say that demand is **perfectly inelastic** in that range.

In the second case (Panel B), the demand is horizontal at some price. Clearly, for an individual consumer, this situation could not occur because it implies that at even an infinitesimally higher price the consumer would buy nothing, whereas at that particular price, the consumer would buy an indeterminately large amount. This situation is not at all an unreasonable description of the demand curve facing a single seller in a perfectly competitive market, such as the wheat market. At the current market price of wheat, an individual farmer could sell all she has. If, however, the farmer held out for a price above market price, it is reasonable that she would not be able to sell any at all because all other farmers' wheat is a perfect substitute for hers, so no one would be willing to buy any of hers at a higher price. In this case, we would say that the demand curve facing a perfectly competitive seller is **perfectly elastic**.

Own-price elasticity of demand is our measure of how sensitive the quantity demanded is to changes in the price of a good or service, but what characteristics of a good or its market might be informative in determining whether demand is highly elastic? Perhaps the most important characteristic is whether there are close substitutes for the good in question. If there are close substitutes for the good, then if its price rises even slightly, a consumer would tend to purchase much less of this good and switch to the substitute, which is now relatively less costly. If there simply are no substitutes, however, then it is likely that the demand is much less elastic. To understand this more fully, consider a consumer's demand for some broadly defined product such as bread. There really are no close substitutes for the broad category bread, which includes all types from French bread to pita bread to tortillas and so on. So, if the price of all bread were to rise, perhaps a consumer would purchase a little less of it each week, but probably not a significantly smaller amount. Now, however, consider that the consumer's demand for a particular baker's specialty bread instead of the category bread as a whole. Surely,

there are closer substitutes for Baker Bob's Whole Wheat Bread with Sesame Seeds than for bread in general. We would expect, then, that the demand for Baker Bob's special loaf is much more elastic than for the entire category of bread. This fact is why the demand faced by an individual wheat farmer is much more elastic than the entire market demand for wheat; there are much closer substitutes for *her* wheat than for wheat *in general.*

In finance, there exists the question of whether the demand for common stock is perfectly elastic. That is, are there perfect substitutes for a firm's common shares? If so, then the demand curve for its shares should be perfectly horizontal. If not, then one would expect a negatively sloped demand for shares. If demand is horizontal, then an increase in demand (owing to some influence other than positive new information regarding the firm's outlook) would not increase the share price. In contrast, a purely mechanical increase in demand would be expected to increase the price if the demand were negatively sloped. One study looked at evidence from 31 stocks whose weights on the Toronto Stock Exchange 300 index were changed, owing purely to fully anticipated technical reasons that apparently had no relationship to new information about those firms.[12] That is, the demand for those shares shifted rightward. The authors found that there was a statistically significant 2.3 percent excess return associated with those shares, a finding consistent with a negatively sloped demand curve for common stock.

In addition to the degree of substitutability, other characteristics tend to be generally predictive of a good's elasticity of demand. These include the portion of the typical budget that is spent on the good, the amount of time that is allowed to respond to the change in price, the extent to which the good is seen as necessary or optional, and so on. In general, if consumers tend to spend a very small portion of their budget on a good, their demand tends to be less elastic than if they spend a very large part of their income. Most people spend only a little on, say, toothpaste each month, so it really doesn't matter whether the price rises 10 percent; they would probably still buy about the same amount. If the price of housing were to rise significantly, however, most households would try to find a way to reduce the quantity they buy, at least in the long run.

This example leads to another characteristic regarding price elasticity. For most goods and services, the long-run demand is much more elastic than the short-run demand. The reason is that if the price were to change for, say, gasoline, we probably would not be able to respond quickly with a significant reduction in the quantity we consume. In the short run, we tend to be locked into modes of transportation, housing and employment location, and so on. The longer the adjustment time, however, the greater the degree to which a household could adjust to the change in price. Hence, for most goods, long-run elasticity of demand is greater than short-run elasticity. Durable goods, however, tend to behave in the opposite way. If the prices of washing machines were to fall, people might react quickly because they have an old machine that they know will need to be replaced fairly soon anyway. So when prices fall, they might decide to go ahead and make the purchase. If the prices of washing machines were to stay low forever, however, it is unlikely that a typical consumer would buy all that many more machines over a lifetime.

Certainly, whether the good or service is seen to be nondiscretionary or discretionary would help determine its sensitivity to a price change. Faced with the same percentage increase in prices, consumers are much more likely to give up their Friday night restaurant meal than

[12]Aditya Kaul, Vikas Mehrotra, and Randall Morck (2000).

they are to cut back significantly on staples in their pantry. The more a good is seen as being necessary, the less elastic its demand is likely to be.

In summary, own-price elasticity of demand is likely to be greater (i.e., more sensitive) for items that have many close substitutes, occupy a large portion of the total budget, are seen to be optional instead of necessary, and have longer adjustment times. Obviously, not all of these characteristics operate in the same direction for all goods, so elasticity is likely to be a complex result of these and other characteristics. In the end, the actual elasticity of demand for a particular good turns out to be an empirical fact that can be learned only from careful observation and, often, sophisticated statistical analysis.

4.2. Own-Price Elasticity of Demand: Impact on Total Expenditure

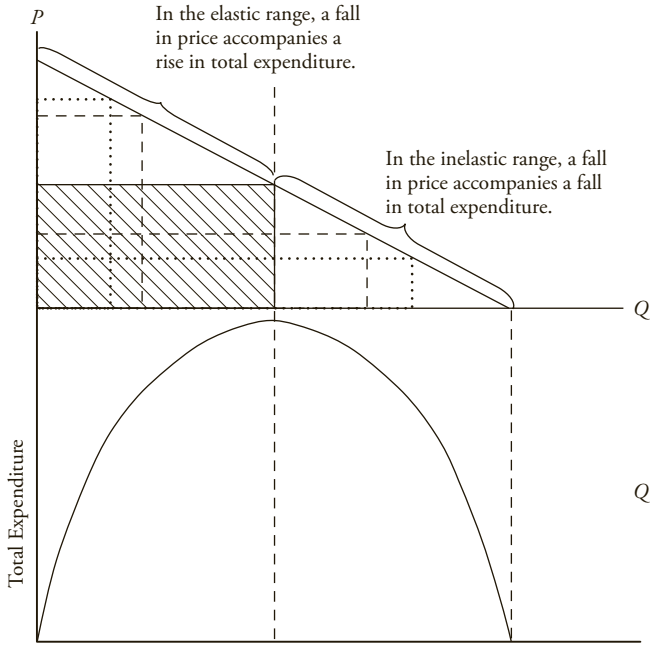
Because of the law of demand, an increase in price is associated with a decrease in the number of units demanded of some good or service. But what can we say about the *total expenditure* on that good? That is, what happens to price times quantity when price falls? Recall that elasticity is defined as the ratio of the percentage change in quantity demanded to the percentage change in price. So if demand is elastic, a decrease in price is associated with a larger percentage rise in quantity demanded. For example, if elasticity were equal to negative 2, then the percentage change in quantity demanded would be twice as large as the percentage change in price. It follows that a 10 percent fall in price would bring about a rise in quantity of greater magnitude, in this case 20 percent. True, each unit of the good has a lower price, but a sufficiently greater number of units are purchased, so total expenditure (price times quantity) would rise as price falls when demand is elastic.

If demand is inelastic, however, a 10 percent fall in price brings about a rise in quantity less than 10 percent in magnitude. Consequently, when demand is inelastic, a fall in price brings about a fall in total expenditure. If elasticity were equal to negative 1 (unitary elasticity), the percentage decrease in price is just offset by an equal and opposite percentage increase in quantity demanded, so total expenditure does not change at all.

In summary, when demand is elastic, price and total expenditure move in *opposite* directions. When demand is inelastic, price and total expenditure move in the *same* direction. When demand is unitary elastic, changes in price are associated with *no change* in total expenditure. This relationship is easy to identify in the case of a linear demand curve. Recall from Exhibit 1-20 that above the midpoint, demand is elastic; and below the midpoint, demand is inelastic. In the upper section of Exhibit 1-22, total expenditure ($P \times Q$) is measured as the area of a rectangle whose base is Q and height is P. Notice that as price falls, the inscribed rectangles at first grow in size but then become their largest at the midpoint of the demand curve. Thereafter, as price continues to fall, total expenditure falls toward zero. In the lower section of Exhibit 1-22, total expenditure is shown for each quantity purchased. Note that it reaches a maximum at the quantity that defines the midpoint, or unit-elastic point, on the demand curve.

It should be noted that the relationships just described hold for any demand curve, so it does not matter whether we are dealing with the demand curve of an individual consumer, the demand curve of the market, or the demand curve facing any given seller. For a market, the total expenditure by buyers becomes the total revenue to sellers in that market. It follows, then, that if market demand is elastic, a fall in price will result in an increase in total revenue to sellers as a whole, and if demand is inelastic, a fall in price will result in a decrease in total revenue to sellers. Clearly, if the demand faced by any given seller were inelastic at the current price, that seller could increase revenue by increasing the price. Moreover, because demand is

EXHIBIT 1-22 Elasticity and Total Expenditure



Note: Figure depicts the relationship among changes in price, changes in quantity, and changes in total expenditure. Maximum total expenditure occurs at the unit-elastic point on a linear demand curve (the crosshatched rectangle).

negatively sloped, the increase in price would decrease total units sold, which would almost certainly decrease total cost. So no one-product seller would ever knowingly choose to set price in the inelastic range of the demand.

4.3. Income Elasticity of Demand: Normal and Inferior Goods

In general, elasticity is simply a measure of how sensitive one variable is to change in the value of another variable. Quantity demanded of a good is a function of not only its own price, but also consumer income. If income changes, the quantity demanded can respond, so the analyst needs to understand the income sensitivity as well as price sensitivity.

Income elasticity of demand is defined as the percentage change in quantity demanded divided by the percentage change in income (I), holding all other things constant, and can be represented as in Equation 1-25.

$$E_I^d = \frac{\%\Delta Q_x^d}{\%\Delta I} = \frac{\frac{\Delta Q_x^d}{Q_x^d}}{\frac{\Delta I}{I}} = \left(\frac{\Delta Q_x^d}{\Delta I}\right)\left(\frac{I}{Q_x^d}\right) \tag{1-25}$$

Note that the structure of this expression is identical to the structure of own-price elasticity in Equation 1-24. Indeed, all elasticity measures that we will examine will have the

same general structure, so essentially if you've seen one, you've seen them all. The only thing that changes is the independent variable of interest. For example, if the income elasticity of demand for some good has a value of 0.8, we would interpret that to mean that whenever income rises by 1 percent, the quantity demanded at each price would rise by 0.8 percent.

Although own-price elasticity of demand will almost always be negative because of the law of demand, *income* elasticity can be negative, positive, or zero. Positive income elasticity simply means that as income rises, quantity demanded also rises, as is characteristic of most consumption goods. We define a good with positive income elasticity as a **normal good**. It is perhaps unfortunate that economists often take perfectly good English words and give them different definitions. When an economist speaks of a normal good, that economist is saying nothing other than that the demand for that particular good rises when income increases and falls when income decreases. Hence, if we find that when income rises people buy more meals at restaurants, then dining out is defined to be a normal good.

For some goods, there is an *inverse* relationship between quantity demanded and consumer income. That is, when people experience a rise in income, they buy absolutely less of some goods, and they buy more of those goods when their income falls. Hence, income elasticity of demand for those goods is negative. By definition, goods with *negative* income elasticity are called **inferior goods**. Again, here the word *inferior* means nothing other than that the income elasticity of demand for that good is observed to be negative. It does not necessarily indicate anything at all about the quality of that good. Typical examples of inferior goods might be rice, potatoes, or less expensive cuts of meat. One study found that income elasticity of demand for beer is slightly negative, whereas income elasticity of demand for wine is significantly positive. An economist would therefore say that beer is inferior whereas wine is normal. Ultimately, whether a good is called inferior or normal is simply a matter of empirical statistical analysis. And a good could be normal for one income group and inferior for another income group. (A BMW 3 Series automobile might very well be normal for a moderate-income group but inferior for a high-income group of consumers. As their respective income levels rose, those in the moderate-income group might purchase more BMWs, starting with the 3 Series, whereas the upper-income group might buy fewer 3 Series as they traded up to a 5 or 7 Series.) Clearly, for some goods and some ranges of income, consumer income might not have an impact on the purchase decision at all. Hence for those goods, income elasticity of demand is zero.

Thinking back to our discussion of the demand curve, recall that we invoked the assumption of "holding all other things constant" when we plotted the relationship between price and quantity demanded. One of the variables we held constant was consumer income. If income were to change, obviously the whole curve would shift one way or the other. For normal goods, a rise in income would shift the entire demand curve upward and to the right, resulting in an increase in demand. If the good were inferior, however, a rise in income would result in a downward and leftward shift in the entire demand curve.

4.4. Cross-Price Elasticity of Demand: Substitutes and Complements

It should be clear by now that any variable on the right-hand side of the demand function can serve as the basis for its own elasticity. Recall that the price of another good might very well have an impact on the demand for a good or service, so we should be able to define an elasticity with respect to the *other* price, as well. That elasticity is called the **cross-price elasticity of demand** and takes on the same structure as own-price elasticity and income elasticity of demand, as represented in Equation 1-26.

$$E_{P_y}^d = \frac{\%\Delta Q_x^d}{\%\Delta P_y} = \frac{\frac{\Delta Q_x^d}{Q_x^d}}{\frac{\Delta P_y}{P_y}} = \left(\frac{\Delta Q_x^d}{\Delta P_y}\right)\left(\frac{P_y}{Q_x^d}\right) \tag{1-26}$$

Note how similar in structure this equation is to own-price elasticity. The only difference is that the subscript on P is now y, indicating the price of some other good, Y, instead of the own-price, X. This cross-price elasticity of demand measures how sensitive the demand for good X is to changes in the price of some other good, Y, holding all other things constant. For some pairs of goods, X and Y, when the price of Y rises, more of good X is demanded. That is, the cross-price elasticity of demand is positive. Those goods are defined to be **substitutes**. Substitutes are defined empirically. If the cross-price elasticity of two goods is positive, they are substitutes, irrespective of whether someone would consider them similar.

This concept is intuitive if you think about two goods that are seen to be close substitutes, perhaps like two brands of beer. When the price of one of your favorite brands of beer rises, what would you do? You would probably buy less of that brand and more of one of the cheaper brands, so the cross-price elasticity of demand would be positive.

Alternatively, two goods whose cross-price elasticity of demand is negative are defined to be **complements**. Typically, these goods would tend to be consumed together as a pair, such as gasoline and automobiles or houses and furniture. When automobile prices fall, we might expect the quantity of autos demanded to rise, and thus we might expect to see a rise in the demand for gasoline. Ultimately, though, whether two goods are substitutes or complements is an empirical question answered solely by observation and statistical analysis. If, when the price of one good rises, the demand for the other good also rises, they are substitutes. If the demand for that other good falls, they are complements. And the result might not immediately resonate with our intuition. For example, grocery stores often put something like coffee on sale in the hope that customers will come in for coffee and end up doing their weekly shopping there as well. In that case, coffee and, say, cabbage could very well empirically turn out to be complements even though we do not normally think of consuming coffee and cabbage together as a pair (i.e., that the price of coffee has a relationship to the sales of cabbage).

For substitute goods, an increase in the price of one good would shift the demand curve for the other good upward and to the right. For complements, however, the impact is in the other direction: When the price of one good rises, the quantity demanded of the other good shifts downward and to the left.

4.5. Calculating Demand Elasticities from Demand Functions

Although the concept of different elasticities of demand is helpful in sorting out the qualitative and directional effects among variables, the analyst will also benefit from having an empirically estimated demand function from which to calculate the magnitudes as well. There is no substitute for actual observation and statistical (regression) analysis to yield insights into the quantitative behavior of a market. (Empirical analysis, however, is outside the scope of this chapter.) To see how an analyst would use such an equation, let us return to our hypothetical market demand function for gasoline in Equation 1-13, duplicated in Equation 1-27:

$$Q_x^d = 8{,}400 - 400P_x + 60I - 10P_y \tag{1-27}$$

As we found when we calculated own-price elasticity of demand earlier, we need to identify where to look by choosing actual values for the independent variables, P_x, I, and P_y.

We choose \$3 for P_x, \$50 (thousands) for I, and \$20 (thousands) for P_y. By inserting these values into the estimated demand function (Equation 1-27), we find that quantity demanded is 10,000 gallons of gasoline per week. We now have everything we need to calculate own-price, income, and cross-price elasticities of demand for our market. Those respective elasticities are expressed in Equations 1-28, 1-29, and 1-30. Each of those expressions has a term denoting the change in quantity divided by the change in each respective variable: $\Delta Q_x/\Delta P_x$, $\Delta Q_x/\Delta I$, and $\Delta Q_x/\Delta P_y$. In each case, those respective terms are given by the coefficients on the variables of interest. Once we recognize this fact, the rest is accomplished simply by inserting values into the elasticity formulas.

$$E^d_{p_x} = \left(\frac{\Delta Q^d_x}{\Delta P_x}\right)\left(\frac{P_x}{Q^d_x}\right) = [-400]\left[\frac{3}{10{,}000}\right] = -0.12 \tag{1-28}$$

$$E^d_I = \left(\frac{\Delta Q^d_x}{\Delta I}\right)\left(\frac{I}{Q^d_x}\right) = [60]\left[\frac{50}{10{,}000}\right] = 0.30 \tag{1-29}$$

$$E^d_{P_y} = \left(\frac{\Delta Q^d_x}{\Delta P_y}\right)\left(\frac{P_y}{Q^d_x}\right) = [-10]\left[\frac{20}{10{,}000}\right] = -0.02 \tag{1-30}$$

In our example, at a price of \$3, the own-price elasticity of demand is -0.12, meaning that a 1 percent increase in the price of gasoline would bring about a decrease in quantity demanded of only 0.12 percent. Because the absolute value of the own-price elasticity is less than 1, we characterize demand as being *inelastic* at that price, so an increase in price would result in an increase in total expenditure on gasoline by consumers in that market. Additionally, the income elasticity of demand is 0.30, meaning that a 1 percent increase in income would bring about an increase of 0.30 percent in the quantity demanded of gasoline. Because that elasticity is positive (but small), we would characterize gasoline as a *normal* good: An increase in income would cause consumers to buy more gasoline. Finally, the cross-price elasticity of demand between gasoline and automobiles is -0.02, meaning that if the price of automobiles rose by 1 percent, the demand for gasoline would fall by 0.02 percent. We would therefore characterize gasoline and automobiles as *complements* because the cross-price elasticity is negative. The magnitude is, however, quite small, so we would conclude that the complementary relationship is quite weak.

EXAMPLE 1-13 Calculating Elasticities from a Given Demand Function

An individual consumer's monthly demand for downloadable e-books is given by the equation $Q^d_{eb} = 2 - 0.4P_{eb} + 0.0005I + 0.15P_{hb}$, where Q^d_{eb} equals the number of e-books demanded each month, I equals the household monthly income, P_{eb} equals the price of e-books, and P_{hb} equals the price of hardbound books. Assume that the price of

e-books is €10.68, household income is €2,300, and the price of hardbound books is €21.40.

1. Determine the value of own-price elasticity of demand for e-books.
2. Determine the income elasticity of demand for e-books.
3. Determine the cross-price elasticity of demand for e-books with respect to the price of hardbound books.

Solution to 1: Recall that own-price elasticity of demand is given by $(\Delta Q_{eb}/\Delta P_{eb})(P_{eb}/Q_{eb})$, and notice from the demand function that $\Delta Q_{eb}/\Delta P_{eb} = -0.4$. Inserting the given variable values into the demand function yields $Q_{eb} = 2.088$. So at a price of €10.68, the own-price elasticity of demand equals $(-0.4)(10.68/2.088) = -2.046$, which is elastic because in absolute value the elasticity coefficient is greater than 1.

Solution to 2: Recall that income elasticity of demand is given by $(\Delta Q_{eb}/\Delta I)(I/Q_{eb})$. Notice from the demand function that $\Delta Q_{eb}/\Delta I = 0.0005$. Inserting in the values for I and Q_{eb} yields income elasticity of $(0.0005)(2{,}300/2.088) = 0.551$, which is positive, so e-books are a normal good.

Solution to 3: Recall that cross-price elasticity of demand is given by $(\Delta Q_{eb}/\Delta P_{hb})(P_{hb}/Q_{eb})$, and notice from the demand function that $\Delta Q_{eb}/\Delta P_{hb} = 0.15$. Inserting in the values for P_{hb} and Q_{eb} yields a cross-price elasticity of demand for e-books of $(0.15)(21.40/2.088) = 1.537$, which is positive, implying that e-books and hardbound books are substitutes.

5. SUMMARY

This chapter has surveyed demand and supply analysis. Because markets (goods markets, factor markets, and capital markets) supply the foundation for today's global economy, an understanding of the demand and supply model is essential for any analyst who hopes to grasp the implications of economic developments on investment values. Among the points made are the following:

- The basic model of markets is the demand and supply model. The demand function represents buyers' behavior and can be depicted (in its inverse demand form) as a negatively sloped demand curve. The supply function represents sellers' behavior and can be depicted (in its inverse supply form) as a positively sloped supply curve. The interaction of buyers and sellers in a market results in equilibrium. Equilibrium exists when the highest price willingly paid by buyers is just equal to the lowest price willingly accepted by sellers.
- Goods markets are the interactions of consumers as buyers and firms as sellers of goods and services produced by firms and bought by households. Factor markets are the interactions of firms as buyers and households as sellers of land, labor, capital, and entrepreneurial

risk-taking ability. Capital markets are used by firms to sell debt or equity to raise long-term capital to finance the production of goods and services.

- Demand and supply curves are drawn on the assumption that everything *except* the price of the good itself is held constant (an assumption known as ceteris paribus or "holding all other things constant"). When something other than price changes, the demand curve or the supply curve will shift relative to the other curve. This shift is referred to as a change in demand or supply, as opposed to quantity demanded or quantity supplied. A new equilibrium generally will be obtained at a different price and a different quantity than before. The market mechanism is the ability of prices to adjust to eliminate any excess demand or supply resulting from a shift in one or the other curve.
- If, at a given price, the quantity demanded exceeds the quantity supplied, there is excess demand and the price will rise. If, at a given price, the quantity supplied exceeds the quantity demanded, there is excess supply and the price will fall.
- Sometimes auctions are used to seek equilibrium prices. Common value auctions sell items that have the same value to all bidders, but bidders can only estimate that value before the auction is completed. Overly optimistic bidders overestimate the true value and end up paying a price greater than that value. This result is known as the winner's curse. Private value auctions sell items that (generally) have a unique subjective value for each bidder. Ascending price auctions use an auctioneer to call out ever-increasing prices until the last, highest bidder ultimately pays his or her bid price and buys the item. Descending price, or Dutch, auctions begin at a very high price and then reduce that price until one bidder is willing to buy at that price. Second price sealed-bid auctions are sometimes used to induce bidders to reveal their true reservation prices in private value auctions. Treasury notes and some other financial instruments are sold using a form of Dutch auction (called a single price auction) in which competitive and noncompetitive bids are arrayed in descending price (increasing yield) order. The winning bidders all pay the same price, but marginal bidders might not be able to fill their entire order at the market-clearing price.
- Markets that work freely can optimize society's welfare, as measured by consumer surplus and producer surplus. Consumer surplus is the difference between the total value to buyers and the total expenditure necessary to purchase a given amount. Producer surplus is the difference between the total revenue received by sellers from selling a given amount and the total variable cost of production of that amount. When equilibrium price is reached, total surplus is maximized.
- Sometimes, government policies interfere with the free working of markets. Examples include price ceilings, price floors, and specific taxes. Whenever the imposition of such a policy alters the free market equilibrium quantity (the quantity that maximizes total surplus), there is a redistribution of surplus between buyers and sellers; but there is also a reduction of total surplus, called deadweight loss. Other influences can result in an imbalance between demand and supply. Search costs are impediments in the ability of willing buyers and willing sellers to meet in a transaction. Brokers can add value if they reduce search costs and match buyers and sellers. In general, anything that improves information about the willingness of buyers and sellers to engage will reduce search costs and add value.
- Economists use a quantitative measure of sensitivity called elasticity. In general, elasticity is the ratio of the percentage change in the dependent variable to the percentage change in the

independent variable of interest. Important specific elasticities include own-price elasticity of demand, income elasticity of demand, and cross-price elasticity of demand.

- Based on algebraic sign and magnitude of the various elasticities, goods can be classified into groups. If own-price elasticity of demand is less than 1 in absolute value, demand is called "inelastic"; it is called "elastic" if own-price elasticity of demand is greater than 1 in absolute value. Goods with positive income elasticity of demand are called normal goods, and those with negative income elasticity of demand are called inferior goods. Two goods with negative cross-price elasticity of demand—a drop in the price of one good causes an increase in demand for the other good—are called complements. Goods with positive cross-price elasticity of demand—a drop in the price of one good causes a decrease in demand for the other—are called substitutes.
- The relationship among own-price elasticity of demand, changes in price, and changes in total expenditure is as follows: If demand is elastic, a reduction in price results in an increase in total expenditure; if demand is inelastic, a reduction in price results in a decrease in total expenditure; if demand is unitary elastic, a change in price leaves total expenditure unchanged.

PRACTICE PROBLEMS[13]

1. Which of the following markets is *most* accurately characterized as a goods market? The market for:
   A. coats.
   B. sales clerks.
   C. cotton farmland.

2. The observation "As a price of a good falls, buyers buy more of it" is *best* known as:
   A. consumer surplus.
   B. the law of demand.
   C. the market mechanism.

3. Two-dimensional demand and supply curves are drawn under which of the following assumptions?
   A. Own price is held constant.
   B. All variables but quantity are held constant.
   C. All variables but own price and quantity are held constant.

4. The slope of a supply curve is *most* often:
   A. zero.
   B. positive.
   C. negative.

5. Assume the following equation:

$$Q_x^s = -4 + \tfrac{1}{2}P_x - 2W$$

[13]These practice problems were written by William Akmentins, CFA (Dallas, Texas, USA).

where Q_x^s is the quantity of good X supplied, P_x is the price of good X, and W is the wage rate paid to laborers. If the wage rate is 11, the vertical intercept on a graph depicting the supply curve is *closest* to:
A. -26.
B. -4.
C. 52.

6. Movement along the demand curve for good X occurs due to a change in:
A. income.
B. the price of good X.
C. the price of a substitute for good X.

The following information relates to Questions 7 through 9.

A producer's supply function is given by the equation:

$$Q_s^s = -55 + 26P_s + 1.3P_a$$

where Q_s^s is the quantity of steel supplied by the market, P_s is the per-unit price of steel, and P_a is the per-unit price of aluminum.

7. If the price of aluminum rises, what happens to the steel producer's supply curve? The supply curve:
A. shifts to the left.
B. shifts to the right.
C. remains unchanged.

8. If the unit price of aluminum is 10, the slope of the supply curve is *closest* to:
A. 0.04.
B. 1.30.
C. 26.00.

9. Assume the supply side of the market consists of exactly five identical sellers. If the unit price of aluminum is 20, which equation is *closest* to the expression for the market inverse supply function?
A. $P_s = 9.6 + 0.04Q_s^s$
B. $P_s = 1.1 + 0.008Q_s^s$
C. $Q_s^s = -145 + 130P_s$

10. Which of the following statements about market equilibrium is *most* accurate?
A. The difference between quantity demanded and quantity supplied is zero.
B. The demand curve is negatively sloped and the supply curve is positively sloped.
C. For any given pair of market demand and supply curves, only one equilibrium point can exist.

11. Which of the following statements *best* characterizes the market mechanism for attaining equilibrium?

A. Excess supply causes prices to fall.
B. Excess demand causes prices to fall.
C. The demand and supply curves shift to reach equilibrium.

12. An auction in which the auctioneer starts at a high price and then lowers the price in increments until there is a willing buyer is *best* called a:
A. Dutch auction.
B. Vickery auction.
C. private-value auction.

13. Which statement is *most likely* to be true in a single price U.S. Treasury bill auction?
A. Only some noncompetitive bids would be filled.
B. Bidders at the highest winning yield may get only a portion of their orders filled.
C. All bidders at a yield higher than the winning bid would get their entire orders filled.

14. The winner's curse in common value auctions is *best* described as the winning bidder paying:
A. more than the value of the asset.
B. a price not equal to one's own bid.
C. more than intended prior to bidding.

15. A wireless phone manufacturer introduced a next-generation phone that received a high level of positive publicity. Despite running several high-speed production assembly lines, the manufacturer is still falling short in meeting demand for the phone nine months after introduction. Which of the following statements is the *most* plausible explanation for the demand/supply imbalance?
A. The phone price is low relative to the equilibrium price.
B. Competitors introduced next-generation phones at a similar price.
C. Consumer incomes grew faster than the manufacturer anticipated.

16. A per-unit tax on items sold that is paid by the seller will *most likely* result in the:
A. supply curve shifting vertically upward.
B. demand curve shifting vertically upward.
C. demand curve shifting vertically downward.

17. Which of the following *most* accurately and completely describes a deadweight loss?
A. A transfer of surplus from one party to another
B. A reduction in either the buyer's or the seller's surplus
C. A reduction in total surplus resulting from market interference

18. If an excise tax is paid by the buyer instead of the seller, which of the following statements is *most likely* to be true?
A. The price paid will be higher than if the seller had paid the tax.
B. The price received will be lower than if the seller had paid the tax.
C. The price received will be the same as if the seller had paid the tax.

19. A quota on an imported good below the market-clearing quantity will *most likely* lead to which of the following effects?
    A. The supply curve shifts upward.
    B. The demand curve shifts upward.
    C. Some of the buyer's surplus transfers to the seller.

20. Assume a market demand function is given by the equation:

$$Q^d = 50 - 0.75P$$

    where Q^d is the quantity demanded and P is the price. If P equals 10, the value of the consumer surplus is *closest* to:
    A. 67.
    B. 1,205.
    C. 1,667.

21. Which of the following *best* describes producer surplus?
    A. Revenue minus variable costs
    B. Revenue minus variable plus fixed costs
    C. The area above the supply curve and beneath the demand curve and to the left of the equilibrium point

22. Assume a market supply function is given by the equation

$$Q_s = -7 + 0.6P$$

    where Q_s is the quantity supplied and P is the price. If P equals 15, the value of the producer surplus is *closest* to:
    A. 3.3.
    B. 41.0.
    C. 67.5.

The following information relates to Questions 23 through 25.

The market demand function for four-year private universities is given by the equation:

$$Q_{pr}^d = 84 - 3.1P_{pr} + 0.8I + 0.9P_{pu}$$

where Q_{pr}^d is the number of applicants to private universities per year in thousands, P_{pr} is the average price of private universities (in thousands of USD), I is the household monthly income (in thousands of USD), and P_{pu} is the average price of public (government-supported) universities (in thousands of USD). Assume that P_{pr} is equal to 38, I is equal to 100, and P_{pu} is equal to 18.

23. The price elasticity of demand for private universities is *closest* to:
    A. −3.1.
    B. −1.9.
    C. 0.6.

24. The income elasticity of demand for private universities is *closest* to:
    A. 0.5.
    B. 0.8.
    C. 1.3.

25. The cross-price elasticity of demand for private universities with respect to the average price of public universities is *closest* to:
    A. 0.3.
    B. 3.1.
    C. 3.9.

26. If the cross-price elasticity between two goods is negative, the two goods are classified as:
    A. normal.
    B. substitutes.
    C. complements.

CHAPTER 2

DEMAND AND SUPPLY ANALYSIS: CONSUMER DEMAND

Richard V. Eastin

Gary L. Arbogast, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Describe consumer choice theory and utility theory.
- Describe the use of indifference curves, opportunity sets, and budget constraints in decision making.
- Calculate and interpret a budget constraint.
- Determine a consumer's equilibrium bundle of goods based on utility analysis.
- Compare substitution and income effects.
- Distinguish between normal goods and inferior goods, and explain Giffen goods and Veblen goods in this context.

1. INTRODUCTION

By now it should be clear that economists are model builders. In the previous chapter, we examined one of their most fundamental models, the model of demand and supply. And as we have seen, models begin with simplifying assumptions and then find the implications that can then be compared to real-world observations as a test of the model's usefulness. In the model of demand and supply, we *assumed* the existence of a demand curve and a supply curve, as well as their respective negative and positive slopes. That simple model yielded some very powerful implications about how markets work, but we can delve even more deeply to explore the underpinnings of demand and supply. In this chapter, we examine the theory of the

consumer as a way of understanding where consumer demand curves originate. In a subsequent chapter, the origins of the supply curve are sought in presenting the theory of the firm.

This chapter is organized as follows: Section 2 describes consumer choice theory in more detail. Section 3 introduces utility theory, a building block of consumer choice theory that provides a quantitative model for a consumer's preferences and tastes. Section 4 surveys budget constraints and opportunity sets. Section 5 covers the determination of the consumer's bundle of goods and how that may change in response to changes in income and prices. Section 6 examines substitution and income effects for different types of goods. A summary and practice problems conclude the chapter.

2. CONSUMER THEORY: FROM PREFERENCES TO DEMAND FUNCTIONS

The introduction to demand and supply analysis in the previous chapter basically assumed that the demand function exists, and focused on understanding its various characteristics and manifestations. In this chapter, we address the foundations of demand and supply analysis and seek to understand the sources of consumer demand through the theory of the consumer, also known as consumer choice theory. **Consumer choice theory** can be defined as the branch of microeconomics that relates consumer demand curves to consumer preferences. Consumer choice theory begins with a fundamental model of how consumer preferences and tastes might be represented. It explores consumers' willingness to trade off between two goods (or two baskets of goods), both of which the consumer finds beneficial. Consumer choice theory then recognizes that to consume a set of goods and services, consumers must purchase them at given market prices and with a limited income. In effect, consumer choice theory first models what the consumer would like to consume, and then it examines what the consumer can consume with limited income. Finally, by superimposing what the consumer would *like* to do onto what the consumer *can* do, we arrive at a model of what the consumer *would* do under various circumstances. Then by changing prices and income, the model develops consumer demand as a logical extension of consumer choice theory.

Although consumer choice theory attempts to model consumers' preferences or tastes, it does not have much to say about *why* consumers have the tastes and preferences they have. It still makes assumptions, but does so at a more fundamental level. Instead of assuming the existence of a demand curve, it derives a demand curve as an implication of assumptions about preferences. Note that economists are not attempting to predict the behavior of any single consumer in any given circumstance. Instead, they are attempting to build a consistent model of aggregate market behavior in the form of a market demand curve.

Once we model the consumer's preferences, we then recognize that consumption is governed not only by preferences but also by the consumer's **budget constraint** (the ability to purchase various combinations of goods and services, given the consumer's income). Putting preference theory together with the budget constraint gives us the demand curve we are seeking. In the following sections, we explore these topics in turn.

3. UTILITY THEORY: MODELING PREFERENCES AND TASTES

At the foundation of consumer behavior theory is the assumption that the consumer knows his or her own tastes and preferences and tends to take rational actions that result in a more

preferred consumption bundle over a less preferred bundle. To build a consistent model of consumer choice, we need to begin with a few assumptions about preferences.

3.1. Axioms of the Theory of Consumer Choice

First, let us be clear about the consumption opportunities over which the consumer is assumed to have preferences. We define a **consumption bundle** or **consumption basket** as a specific combination of the goods and services that the consumer would like to consume. We could almost literally conceive of a basket containing a given amount of, say, shoes, pizza, medical care, theater tickets, piano lessons, and all the other things that a consumer might enjoy consuming. Each of those goods and services can be represented in a given basket by a nonnegative quantity, respectively, of all the possible goods and services. Any given basket could have zero of one or more of those goods. A distinctly different consumption bundle would contain all of the same goods but in different quantities, again allowing for the possibility of a zero quantity of one or more of the goods. For example, bundle A might have the same amount of all but one of the goods and services as bundle B but a different amount of that one. Bundles A and B would be considered two distinct bundles.

Given this understanding of consumption bundles, the first assumption we make about a given consumer's preferences is simply that she is able to make a comparison between any two possible bundles. That is, given bundles A and B, she must be able to say either that she prefers A to B, or she prefers B to A, or she is *indifferent* between the two. This is the assumption of **complete preferences** (also known as the axiom of completeness), and although it does not appear to be a particularly strong assumption, it is not trivial, either. It rules out the possibility that the consumer could just say, "I recognize that the two bundles are different, but in fact they are *so* different that I simply cannot compare them at all." A loving father might very well say that about his two children. In effect, the father neither prefers one to the other nor is, in any meaningful sense, indifferent between the two. The assumption of complete preferences cannot accommodate such a response.

Second, we assume that when comparing any three distinct bundles, A, B, and C, if A is preferred to B, and simultaneously B is preferred to C, then it must be true that A is preferred to C. This assumption is referred to as the assumption of **transitive preferences**, and it is assumed to hold for indifference as well as for strict preference. This is a somewhat stronger assumption because it is essentially an assumption of rationality. We would say that if a consumer prefers a skiing holiday to a diving holiday and a diving holiday to a backpacking holiday and at the same time prefers a backpacking holiday to a skiing holiday, then he is acting irrationally. Transitivity rules out this kind of inconsistency. If you have studied psychology, however, you will no doubt have seen experiments that show subjects violating this assumption, especially in cases of many complex options being offered to them.

When we state these axioms, we are not saying that we believe them actually to be true in every instance, but we assume them for the sake of building a model. A model is a simplification of the real-world phenomena we are trying to understand. Necessarily, axioms must be at some level inaccurate and incomplete representations of the phenomena we are trying to model. If that were not the case, the model would not be a simplification; it would be a reflection of the complex system we are attempting to model and thus would not help our understanding very much.

Finally, we usually assume that in at least one of the goods, the consumer could never have so much that she would refuse any more, even if it were free. This assumption is sometimes referred to as the "more is better" assumption or the assumption of **nonsatiation**.

Clearly, for some things, more *is* worse, such as air pollution or trash. In those cases, the *good* is then the *removal* of that *bad*, so we can usually reframe our model to accommodate the nonsatiation assumption. In particular, when we later discuss the concept of risk for an investor, we will recognize that for many, more risk is worse than less risk, all else being equal. In that analysis, we will model the willingness of the investor to trade off between increased investment returns and increased certainty, which is the absence of risk.

EXAMPLE 2-1 Axioms Concerning Preferences

Helen Smith enjoys, among other things, eating sausages. She also enjoys reading Marcel Proust. Smith is confronted with two baskets: basket A, which contains several other goods and a package of sausages, and B, which contains identical quantities of the other goods as basket A, but instead of the sausages, it contains a book by Proust. When asked which basket she prefers, she replies, "I like them both, but sausages and a book by Proust are *so* different that I simply cannot compare the two baskets." Determine whether Smith is obeying all the axioms of preference theory.

Solution: Smith is violating the assumption of complete preferences. This assumption states that a consumer must be able to compare any two baskets of goods, either preferring one to the other or being indifferent between the two. If she complies with this assumption, she must be able to compare these two baskets of goods.

3.2. Representing the Preference of a Consumer: The Utility Function

Armed with the assumptions of completeness, transitivity, and nonsatiation, we ask whether there might be a way for a given consumer to represent his own preferences in a consistent manner. Let us consider presenting him with all possible bundles of all the possible goods and services he could consider. Now suppose we give him paper and pencil and ask him to assign a number to each of the bundles. (The assumption of completeness ensures that he, in fact, could do that.) All he must do is write a number on a paper and lay it on each of the bundles. The only restrictions are these: Comparing any two bundles, if he prefers one to the other, he must assign a higher number to the bundle he prefers. And if he is indifferent between them, he must assign the same number to both. Other than that, he is free to begin with any number he wants for the first bundle he considers. In this way, he is simply ordering the bundles according to his preferences over them.

Of course, each of these possible bundles has a specific quantity of each of the goods and services. So, we have two sets of numbers. One set consists of the pieces of paper he has laid on the bundles. The other is the set of numerical quantities of the goods that are contained in each of the respective bundles. Under *reasonable assumptions* (it is not necessary for us to delve into the definition of reasonable assumptions at this level), it is possible to come up with a rule that translates the quantities of goods in each basket into the number that our consumer has assigned to each basket. That assignment rule is called the **utility function** of that particular consumer. The single task of that utility function is to translate each basket of goods and services into a number that rank orders the baskets according to our particular consumer's

preferences. The number itself is referred to as the utility of that basket and is measured in **utils**, which are just quantities of happiness, or well-being, or whatever comes to mind such that more of it is better than less of it.

In general, we can represent the utility function as:

$$U = f(Q_{x_1}, Q_{x_2}, \ldots, Q_{x_n}) \tag{2-1}$$

where the Qs are the quantities of each of the respective goods and services in the bundles. In the case of two goods—say, ounces of wine (W) and slices of bread (B)—a utility function might be simply:

$$U = f(W, B) = WB \tag{2-2}$$

or the product of the number of ounces of wine and the number of slices of bread. The utility of a bundle containing four ounces of wine along with two slices of bread would equal eight utils, and it would rank lower than a bundle containing three ounces of wine along with three slices of bread, which would yield nine utils.

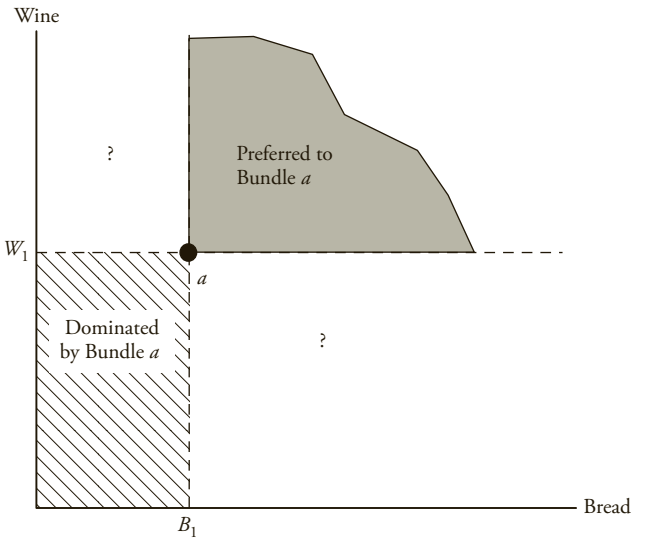
The important point to note is that the utility function is just a ranking of bundles of goods. If someone were to replace all those pieces of paper with new numbers that maintained the same ranking, then the new set of numbers would be just as useful a utility function as the first in describing our consumer's preferences. This characteristic of utility functions is called an *ordinal*, as contrasted to a *cardinal*, ranking. Ordinal rankings are weaker measures than cardinal rankings because they do not allow the calculation and ranking of the *differences* between bundles.

3.3. Indifference Curves: The Graphical Portrayal of the Utility Function

It will be convenient for us to represent our consumer's preferences graphically, not just mathematically. To that end, we introduce the concept of an **indifference curve**, which represents all the combinations of two goods such that the consumer is entirely indifferent between them. This is how we construct such a curve: Consider bundles that contain only two goods so that we can use a two-dimensional graph to represent them—as in Exhibit 2-1, where a particular bundle containing W_a ounces of wine along with B_a slices of bread is represented as a single point, a. The assumption of nonsatiation (more is always better) ensures that all bundles lying directly above, directly to the right of, or both above and to the right of (more wine and more bread) point a must be preferred to bundle a. That set of bundles is called the "preferred to bundle a" set. Correspondingly, all the bundles that lie directly below, to the left of, and both below and to the left of bundle a must yield less utility and therefore would be called the "dominated by bundle a" set.

To determine our consumer's preferences, suppose we present a choice between bundle a and some bundle a', which contains more bread but less wine than a. Nonsatiation is not helpful to us in this case, so we need to ask the consumer which he prefers. If he strictly prefers a', then we would remove a little bread and ask again. If he strictly prefers a, then we would add a little bread, and so on. Finally, after a series of adjustments, we could find just the right combination of bread and wine such that the new bundle a' would be equally satisfying to our consumer as bundle a. That is to say, our consumer would be indifferent between consuming bundle a and consuming bundle a'. We would then choose a bundle, say a'', that contains more wine and less bread than bundle a, and we would again adjust the goods such that the

EXHIBIT 2-1 Showing Preferences Graphically



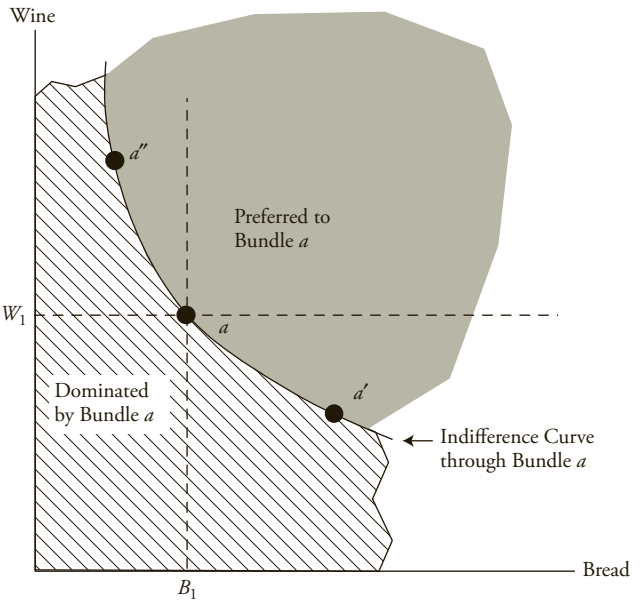
Note: A given bundle of two goods is represented as a single point, *a*, in the two-dimensional graph. Nonsatiation allows us to rank order many, but not all, other bundles relative to *a*, leaving some questions unanswered.

consumer is once again indifferent between bundle *a* and bundle *a*″. By continuing to choose bundles and make adjustments, it would be possible to identify all possible bundles such that the consumer is just indifferent between each of them and bundle *a*. Such a set of points is represented in Exhibit 2-2, where the indifference curve through point *a* represents that set of bundles. Notice that the "preferred to bundle *a*" set has expanded to include all bundles that lie in the region above and to the right of the indifference curve. Correspondingly, the "dominated by bundle *a*" set has expanded to include all bundles that lie in the region below and to the left of the indifference curve.

The indifference curve represents our consumer's unique preferences over the two goods, wine and bread. Its negative slope simply represents that both wine and bread are seen as "good" to this consumer; in order to maintain indifference, a decrease in the quantity of wine must be compensated for by an increase in the quantity of bread. Its curvature tells us something about the strength of the consumer's willingness to trade off one good for the other. The indifference curve in Exhibit 2-2 is characteristically drawn to be *convex* when viewed from the origin. This indicates that the willingness to give up wine to obtain a little more bread diminishes the more bread and the less wine the bundle contains.

We capture this willingness to give up one good to obtain a little more of the other in the phrase **marginal rate of substitution** of bread for wine, MRS_{BW}. The MRS_{BW} is the rate at which the consumer is willing to give up wine to obtain a small increment of bread, holding utility constant (i.e., movement along an indifference curve). Notice that the convexity implies that at a bundle like *a*″, which contains rather a lot of wine and not much bread, the consumer would be willing to give up a considerable amount of wine in exchange for just a little more bread. (The slope of the indifference curve is quite steep at that point.) However, at a point like *a*′, which contains considerably more bread but less wine than *a*″, the consumer is not

EXHIBIT 2-2 An Indifference Curve



Note: An indifference curve shows all combinations of two goods such that the consumer is indifferent between them.

ready to sacrifice nearly as much wine to obtain a little more bread. This suggests that the value being placed on bread, in terms of the amount of wine the consumer is willing to give up for bread, diminishes the more bread and less wine the consumer has. It follows that the MRS_{BW} is the negative of the slope of the tangent to the indifference curve at any given bundle. If, at some point, the slope of the indifference curve had value -2.5, it means that, starting at that particular bundle, our consumer would be willing to sacrifice wine to obtain bread at the rate of 2.5 ounces of wine per slice of bread. Because of the convexity assumption—that MRS_{BW} must diminish as our consumer moves toward more bread and less wine—the MRS_{BW} is continuously changing as he moves along the indifference curve.

EXAMPLE 2-2 Understanding the Marginal Rate of Substitution

Tom Warren currently has 50 blueberries and 20 peanuts. His marginal rate of substitution of peanuts for blueberries, MRS_{pb}, equals 4, and his indifference curves are strictly convex.

1. Determine whether Warren would be willing to trade at the rate of three of his blueberries in exchange for one more peanut.

2. Suppose that Warren is indifferent between his current bundle and one containing 40 blueberries and 25 peanuts. Describe Warren's MRS_{pb} evaluated at the new bundle.

Solution to 1: $MRS_{pb} = 4$ means that Warren would be willing to give up four blueberries for one peanut at that point. He clearly would be willing to give up blueberries at a rate less than that, namely, three to one.

Solution to 2: The new bundle has more peanuts and fewer blueberries than the original one, and Warren is indifferent between the two, meaning that both bundles lie on the same indifference curve, where blueberries are plotted on the vertical axis and peanuts on the horizontal axis. Because his indifference curves are strictly convex and the new bundle lies below and to the right of his old bundle, his MRS_{pb} must be less than 4. That is to say, his indifference curve at the new point must be less steep than at the original bundle.

3.4. Indifference Curve Maps

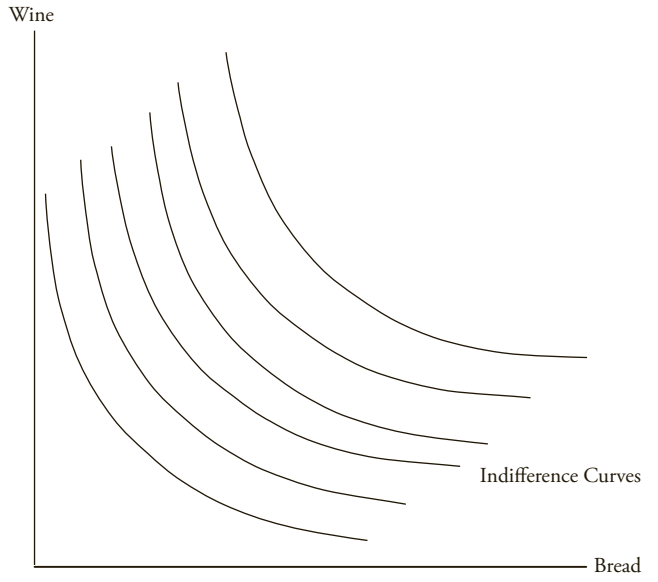
There was nothing special about our initial choice of bundle *a* as a starting point for the indifference curve. We could have begun with a bundle containing more of both goods. In that case, we could have gone through the same process of trial and error, and we would have ended up with another indifference curve, this one passing through the new point and lying above and to the right of the first one. Indeed, we could construct any number of indifference curves in the same manner simply by starting at a different initial bundle. The result is an entire family of indifference curves, called an **indifference curve map**, and it represents our consumer's entire utility function. The word *map* is appropriate because the entire set of indifference curves comprises a contour map of this consumer's utility function. Each contour, or indifference curve, is a set of points in which each point shares a common level of utility with the others. Moving upward and to the right from one indifference curve to the next represents an increase in utility, and moving down and to the left represents a decrease. The map could look like that in Exhibit 2-3.

Because of the completeness assumption, there will be one indifference curve passing through every point in the set. Because of the transitivity assumption, no two indifference curves for a given consumer can ever cross. Exhibit 2-4 shows why. If bundle *a* and bundle *b* lie on the same indifference curve, the consumer must be indifferent between the two. If *a* and *c* lie on the same indifference curve, the consumer must be indifferent between these two bundles as well. But because bundle *c* contains more of both wine and bread than bundle *b*, the consumer must prefer *c* to *b*, which violates transitivity of preferences. So we see that indifference curves will generally be strictly convex and negatively sloped, and they cannot cross. These are the only restrictions we place on indifference curve maps.

3.5. Gains from Voluntary Exchange: Creating Wealth through Trade

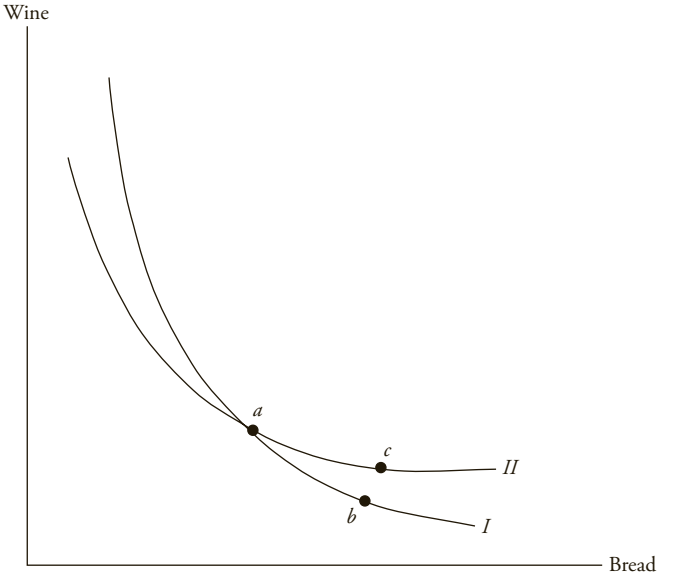
There is no requirement that all consumers have the same preferences. Take the case of Helen Smith and Tom Warren. The indifference curves for Smith will likely be different from Warren's. And although *for any given individual* two indifference curves cannot cross, there is no reason why two indifference curves for two different consumers cannot intersect. Consider Exhibit 2-5, in which we observe an indifference curve for Smith and one for Warren.

EXHIBIT 2-3 An Indifference Curve Map



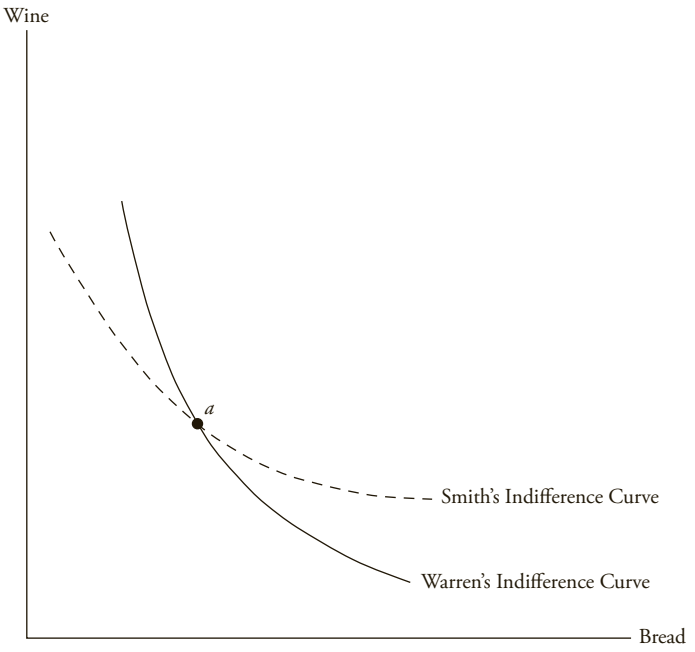
Note: The indifference curve map represents the consumer's utility function. Any curve above and to the right represents a higher level of utility.

EXHIBIT 2-4 Why One Person's Indifference Curves Cannot Cross



Note: Two indifference curves for a given individual cannot cross because the transitivity assumption would be violated.

EXHIBIT 2-5 Two Consumers with Different Preferences



Note: When two consumers have different preferences, they will have different marginal rates of substitution when evaluated at identical bundles. Here, Warren has a relatively strong preference for bread because he is willing to give up more wine for another slice of bread than is Smith.

Suppose they are initially endowed with identical bundles, represented by *a*. They each have exactly identical quantities of bread and wine. Note, however, that because their indifference curves intersect at that point, their slopes are different. Warren's indifference curve is steeper at point *a* than is Smith's. This means that Warren's MRS_{BW} is greater than Smith's MRS_{BW}. That is to say, Warren is willing, at that point, to give up more wine for an additional slice of bread than Smith is. That also means that Smith is willing to give up more bread for an additional ounce of wine than Warren is. Therefore, we observe that Warren has a relatively strong preference for bread compared to Smith, and Smith has a stronger preference for wine than Warren has.

Suppose that the slope of Warren's indifference curve at point *a* is equal to -2, and the slope of Smith's indifference curve at point *a* is equal to $-\frac{1}{2}$. Warren is willing to give up two ounces of wine for one slice of bread, and Smith is willing to give up only a half ounce of wine for one slice of bread. But that means she would be willing to give up 2 slices of bread for 1 ounce of wine. What would happen if Warren and Smith are allowed to exchange bread for wine? Suppose they are allowed to exchange at the ratio of one ounce of wine for one slice of bread. Would they both agree to an exchange at that ratio? Yes. Since Warren is willing to give up two ounces of wine for a slice of bread, he would certainly be willing to give up only one ounce of wine for one slice of bread. Correspondingly, since Smith is willing to give up two slices of bread for one ounce of wine, she would certainly be willing to give up only one slice of bread for one ounce of wine. If they actually made such a trade at the one-to-one ratio, then

Smith would end up with more wine and less bread than she started with, and Warren would end up with more bread and less wine than he started with.

We could say that Warren is better off by the value to him of one ounce of wine because he was *willing* to give up two ounces but only *had to* give up one ounce for his slice of bread. What about Smith? She is better off by the value to her of one slice of bread because she was *willing* to give up two slices of bread for her one additional ounce of wine but only *had to* give up one slice. Both Smith and Warren are better off after they trade. There is no more bread or wine than when they began, but there is greater *wealth* because both are better off than before they traded with each other. Both Smith and Warren ended on higher indifference curves than when they began.

As Smith gives up slices of bread for more ounces of wine, her MRS_{BW} increases; her indifference curve becomes steeper. Simultaneously, as Warren gives up ounces of wine for more slices of bread, his MRS_{BW} decreases; his indifference curve becomes less steep. Eventually, if they continue to trade, their MRSs will reach equality and there will be no further gains to be achieved from additional exchange. Initially, it was the differences in their willingness to trade one good for the other that made trading beneficial to both. But if they trade to a pair of bundles at which their MRSs are equal, then trading will cease.

EXAMPLE 2-3 Understanding Voluntary Exchange

Helen Smith and Tom Warren have identical baskets containing books (B) and compact discs (D). Smith's MRS_{BD} equals 0.8 (i.e., she is willing to give up 0.8 disc for one book), and Warren's MRS_{BD} equals 1.25.

1. Determine whether Warren would accept the trade of one of Smith's discs in exchange for one of his books.
2. State and justify whether Smith or Warren has a stronger preference for books.
3. Determine whether Smith or Warren would end up with more discs than he or she had to begin with, assuming they were allowed to exchange at the rate of one book for one disc. Justify your answer.

Solution to 1: Warren's MRS_{BD} equals 1.25, meaning that he is willing to give up 1.25 discs for one more book. Another way to say this is that Warren requires at least 1.25 discs to compensate him for giving up one book. Because Smith offers only one disc, Warren will not accept the offer. (Of course, Smith would not voluntarily give up one disc for one of Warren's books. Her MRS_{BD} is only 0.8, meaning that she would be willing to give up, at most, 0.8 disc for a book; so she would not have offered one disc for a book anyway.)

Solution to 2: Because Warren is willing to give up 1.25 discs for a book and Smith is willing to give up only 0.8 disc for a book, Warren has a stronger preference for books.

Solution to 3: Smith would have more discs than she originally had. Because Smith has a stronger preference for discs and Warren has a stronger preference for books, Smith would trade books for discs and so would end up with more discs.

4. THE OPPORTUNITY SET: CONSUMPTION, PRODUCTION, AND INVESTMENT CHOICE

So far, we have examined the trade-offs that economic actors (e.g., consumers, companies, investors) are *willing* to make. In this section, we recognize that circumstances almost always impose constraints on the trade-offs that these actors are *able* to make. In other words, we need to explore how to model the constraints on behavior that are imposed by the fact that we live in a world of scarcity: There is simply not enough of everything to satisfy the needs and desires of everyone at a given time. Consumers must generally purchase goods and services with their limited incomes and at given market prices. Companies, too, must divide their limited input resources in order to produce different products. Investors are not able to choose *both* high returns *and* low risk simultaneously. Choices must be made, and here we examine how to represent the set of choices from which to choose.

4.1. The Budget Constraint

Previously, we examined what would happen if Warren and Smith were each given an endowment of bread and wine and were allowed to exchange at some predetermined ratio. Although that circumstance is possible, a more realistic situation would be if Warren or Smith had a given income with which to purchase bread and wine at fixed market prices. Let Warren's income be given by I, the price he must pay for a slice of bread be P_B, and the price he must pay for an ounce of wine be P_W. Warren has freedom to spend his income any way he chooses, as long as the expenditure on bread plus the expenditure on wine does not exceed his income per time period. We can represent this **income constraint** (or budget constraint) with the following expression:

$$P_B Q_B + P_W Q_W \leq I \tag{2-3}$$

This expression simply constrains Warren to spend, in total, no more than his income. At this stage of our analysis, we are assuming a one-period model. In effect, then, Warren has no reason *not* to spend all of his income. The weak inequality becomes a strict equality, as shown in Equation 2-4, because there would be no reason for Warren to save any of his income if there is no tomorrow.

$$P_B Q_B + P_W Q_W = I \tag{2-4}$$

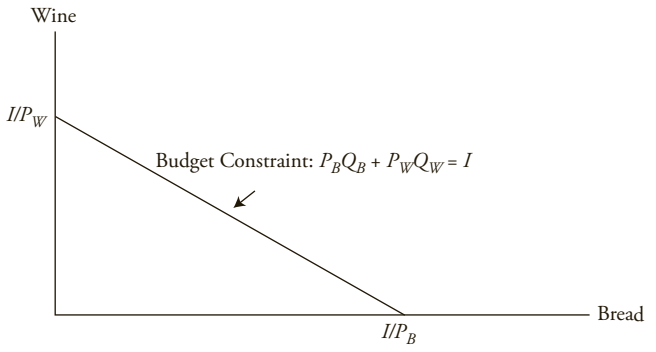
From this equation, we see that if Warren were to spend all of his income only on bread, he could buy I/P_B slices of bread. Or if he were to confine his expenditure to wine alone, he could buy I/P_W ounces of wine. Alternatively, he could spread his income across bread and wine expenditures any way he chooses. Graphically, then, his budget constraint would appear as in Exhibit 2-6.

A simple algebraic manipulation of Equation 2-4 yields the budget constraint in the form of an intercept and slope:

$$Q_W = \frac{I}{P_W} - \frac{P_B}{P_W} Q_B \tag{2-5}$$

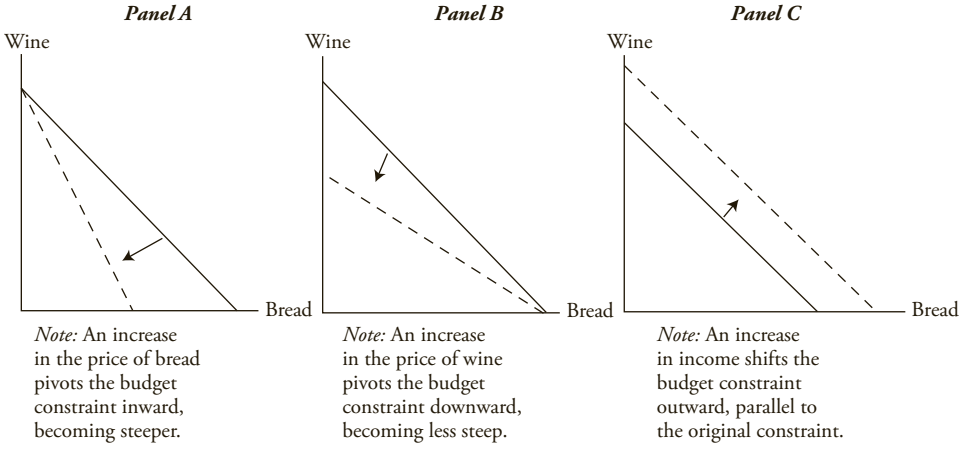
Notice that the slope of the budget constraint is equal to $-P_B/P_W$, and it shows the amount of wine that Warren would have to give up if he were to purchase another slice of

EXHIBIT 2-6 The Budget Constraint



Note: The budget constraint shows all the combinations of bread and wine that the consumer could purchase with a fixed amount of income, I, paying prices P_B and P_W, respectively.

EXHIBIT 2-7 Changing Prices and Income



bread. If the price of bread were to rise, the budget constraint would become steeper, pivoting through the vertical intercept. Alternatively, if the price of wine were to rise, the budget constraint would become less steep, pivoting downward through the horizontal intercept. If income were to rise, the entire budget constraint would shift outward, parallel to the original constraint, as shown in Exhibit 2-7.

As a specific example of a budget constraint, suppose Smith has $60 to spend on bread and wine per month, the price of a slice of bread is $0.50, and the price of an ounce of wine is $0.75. If she spent all of her income on bread, she could buy 120 slices of bread. Or she could buy up to 80 ounces of wine if she chose to buy no bread. Obviously, she can spend half her income on each good, in which case she could buy 60 slices of bread and 40 ounces of wine. The entire set of bundles that Smith could buy with her $60 budget is shown in Exhibit 2-8.

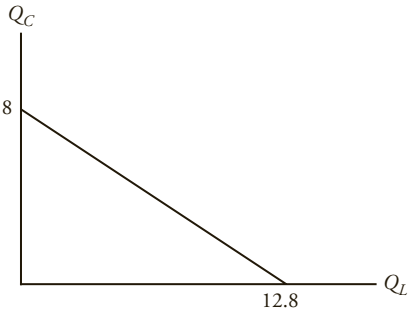
EXAMPLE 2-4 The Budget Constraint

Nigel's Pub has a total budget of £128 per week to spend on cod and lamb. The price of cod is £16 per kilogram, and the price of lamb is £10 per kilogram.

1. Calculate Nigel's budget constraint.
2. Construct a diagram of Nigel's budget constraint.
3. Determine the slope of Nigel's budget constraint.

Solution to 1: The budget constraint is simply that the sum of the expenditure on cod plus the expenditure on lamb be equal to Nigel's budget: $128 = 16Q_C + 10Q_L$. Rearranging, it can also be written in intercept slope form: $Q_C = 128/P_C - (P_L/P_C)$ $Q_L = 8 - 0.625Q_L$.

Solution to 2: We can choose to measure either commodity on the vertical axis, so we arbitrarily choose cod. Note that if Nigel spends his entire budget on cod, he could buy 8 kg. However, if he chooses to spend the entire budget on lamb, he could buy 12.8 kg. Of course, he could spread his £128 between the two goods in any proportions he chooses, so the budget constraint is drawn as follows:

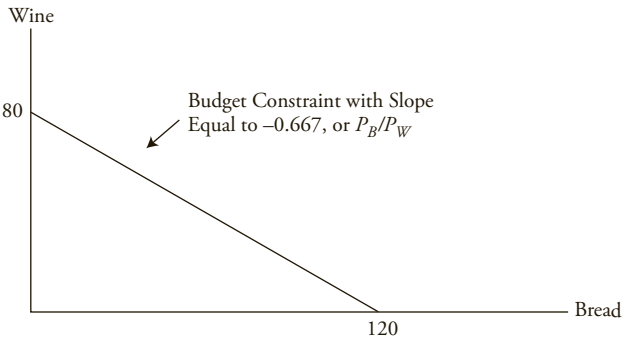


Solution to 3: With quantity of cod measured on the vertical axis, the slope is equal to $-(P_L/P_C) = -10/16 = -0.625$. *Note:* If we had chosen to measure quantity of lamb on the vertical axis, the slope would be inverted: $-(P_C/P_L) = -1.6$.

4.2. The Production Opportunity Set

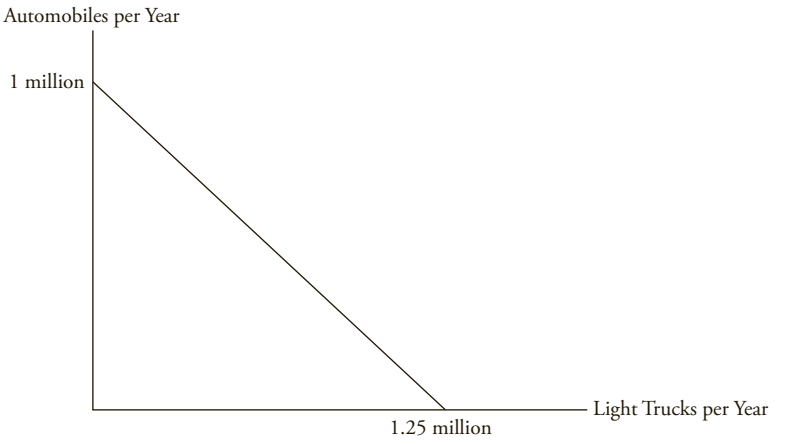
Companies face constraints on their production opportunities, just as consumers face limits on the bundles of goods that they can consume. Consider a company that produces two products using the same production capacity. For example, an automobile company might use the same factory to produce either automobiles or light trucks. If so, then the company is constrained by the limited capacity to produce vehicles. If it produces more trucks, it must reduce its production of automobiles; likewise, if it produces more automobiles, it must produce fewer trucks. The company's **production opportunity frontier** shows the maximum number of units of one good it can produce, for any given number of the other good that it chooses to manufacture. Such a frontier for the vehicle company might look something like that in Exhibit 2-9.

EXHIBIT 2-8 A Specific Example of a Budget Constraint



Note: This exhibit shows Smith's budget constraint if she has an income of $60 and must pay $0.50 per slice of bread and $0.75 per ounce of wine.

EXHIBIT 2-9 The Production Opportunity Frontier



Note: The production opportunity frontier for a vehicle manufacturer shows the maximum number of autos for any given level of truck production. In this example, the opportunity cost of a truck is 0.8 autos.

There are two important things to notice about this example. First, if the company devoted its entire production facility to the manufacture of automobiles, it could produce 1 million in a year. Alternatively, if it devoted its entire plant to trucks, it could produce 1.25 million a year. Of course, it could devote only part of the year's production to trucks, in which case it could produce automobiles during the remainder of the year. In this simple example, for every additional truck the company chooses to make, it would have to produce 0.8 fewer cars. That is, the **opportunity cost** of a truck is 0.8 cars, or the opportunity cost of a car is 1.25 trucks. The opportunity cost of trucks is the negative of the slope of the production opportunity frontier: 1/1.25. And of course, the opportunity cost of an automobile is the inverse of that ratio, or 1.25.

The other thing to notice about this exhibit is that it assumes that the opportunity cost of a truck is independent of how many trucks (and cars) the company produces. The production

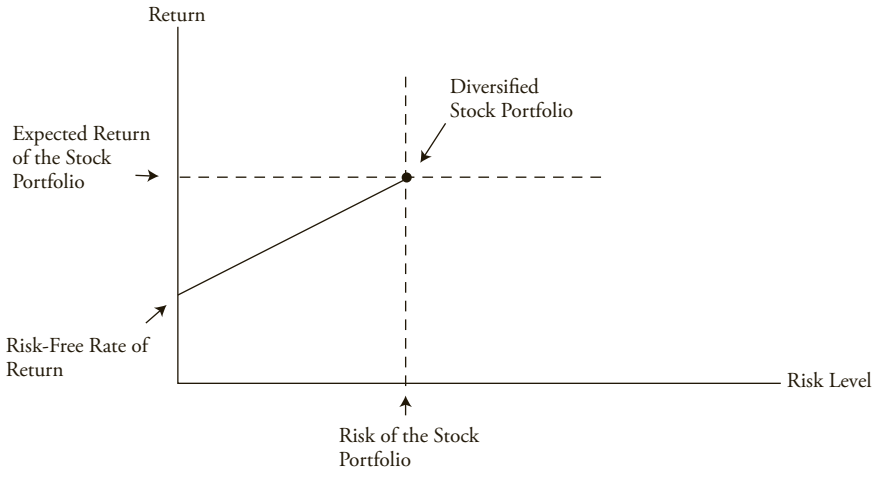
opportunity frontier is linear, with a constant slope. Perhaps a more realistic example would be to increase marginal opportunity cost. As more and more trucks are produced, fewer inputs that are particularly well suited to producing trucks could be transferred to assist in their manufacture, causing the cost of trucks (in terms of cars) to rise as more trucks are produced. In this event, the production opportunity frontier would become steeper as the company moved its production point away from cars and toward more trucks, resulting in a frontier that would be concave as viewed from the origin.

4.3. The Investment Opportunity Set

The investment opportunity set is examined in detail in chapters on investments, but it is appropriate to examine it briefly here because we are learning about constraints on behavior. Consider possible investments in which one option might be to invest in an essentially risk-free asset, such as a U.S. Treasury bill. There is virtually no possibility that the U.S. government would default on a 90-day obligation to pay back an investor's purchase price, plus interest. Alternatively, an investor could put her money into a broadly diversified index of common shares. This investment will necessarily be more risky because of the fact that share prices fluctuate. If investors inherently find risk distasteful, then they will be reluctant to invest in a risky asset unless they expect to receive, on average, a higher rate of return. Hence, it is reasonable to look for a broadly diversified index of common shares to have an expected return exceeding that of the risk-free asset, or else no one would hold that portfolio.

Our hypothetical investor could choose to put some of her funds in the risk-free asset and the rest in the common shares index. For each additional dollar invested in the common shares index, she can expect to receive a higher return, though not with certainty; so, she is exposing herself to more risk in the pursuit of a higher return. We can structure her investment opportunities as a frontier that shows the highest expected return consistent with any given level of risk, as shown in Exhibit 2-10. The investor's choice of a portfolio on the frontier will depend on her level of risk aversion.

EXHIBIT 2-10 The Investment Opportunity Frontier



Note: The investment opportunity frontier shows that as the investor chooses to invest a greater proportion of assets in the market portfolio, she can expect a higher return but also higher risk.

5. CONSUMER EQUILIBRIUM: MAXIMIZING UTILITY SUBJECT TO THE BUDGET CONSTRAINT

It would be wonderful if we could all consume as much of everything as we wanted, but unfortunately, most of us are constrained by income and prices. We now superimpose the budget constraint onto the preference map to model the actual choice of our consumer. This is a constrained (by the resources available to pay for consumption) optimization problem that every consumer must solve: choose the bundle of goods and services that gets us as high on our ranking as possible, while not exceeding our budget.

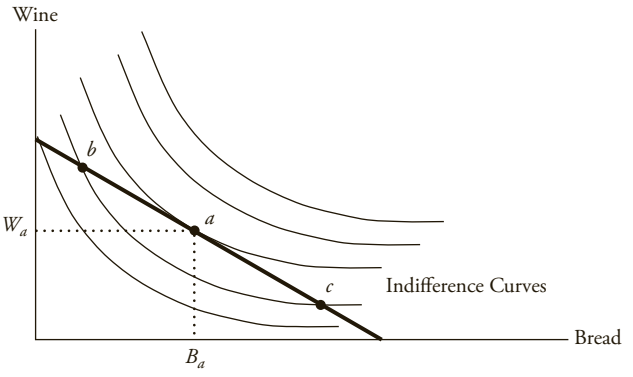
5.1. Determining the Consumer's Equilibrium Bundle of Goods

In general, the consumer's constrained optimization problem consists of maximizing utility, subject to the budget constraint. If, for simplicity, we assume there are only two goods, wine and bread, then the problem appears graphically as in Exhibit 2-11.

The consumer desires to reach the indifference curve that is farthest from the origin while not violating the budget constraint. In this case, that pursuit ends at point *a*, where the consumer is purchasing W_a ounces of wine along with B_a slices of bread per month. It is important to note that this equilibrium point represents the tangency between the highest indifference curve and the budget constraint. At a tangency point, the two curves have the same slope, meaning that the MRS_{BW} must be equal to the price ratio, P_B/P_W. Recall that the marginal rate of substitution is the rate at which the consumer is just *willing* to sacrifice wine for bread. Additionally, the price ratio is the rate at which the consumer *must* sacrifice wine for another slice of bread. So, at equilibrium, the consumer is just willing to pay the opportunity cost that must be paid.

In contrast, consider another affordable bundle represented by point *b*. Certainly, the consumer is able to purchase that bundle because it lies on her budget constraint. However, the MRS_{BW} at that point is greater than the price ratio, meaning that she is *willing* to give up wine to obtain bread at a rate greater than she *must*. Hence, she will be better off moving downward along the budget constraint until she reaches the tangent point at *a*. In effect, the

EXHIBIT 2-11 Consumer Equilibrium



Note: Consumer equilibrium is achieved at point *a*, where the highest indifference curve is attained while not violating the budget constraint.

consumer is willing to pay a higher price than she must for each additional unit of bread until she reaches B_a. For all of the units that she consumes up to B_a, we could say that the consumer is receiving consumer surplus, a concept we visited earlier when discussing the demand curve. Importantly, the consumer would not purchase slices of bread beyond B_a at these prices because at a point like *c*, the marginal rate of substitution is less than the price ratio—meaning that the price for that additional unit is above her willingness to pay. Even though the consumer could afford bundle *c*, it would not be the best use of her income.

EXAMPLE 2-5 Consumer Equilibrium

Currently, a consumer is buying both sorbet and gelato each week. His MRS_{GS}, or marginal rate of substitution of gelato (*G*) for sorbet (*S*), equals 0.75. The price of gelato is €1 per scoop, and the price of sorbet is €1.25 per scoop.

1. Determine whether the consumer is currently optimizing his budget over these two desserts. Justify your answer.
2. Explain whether the consumer should buy more sorbet or more gelato, given that he is not currently optimizing his budget.

Solution to 1: In this example, the condition for consumer equilibrium is $MRS_{GS} = P_G/P_S$. Because $P_G/P_S = 0.8$ and $MRS_{GS} = 0.75$, the consumer is clearly not allocating his budget in a way that maximizes his utility, subject to his budget constraint.

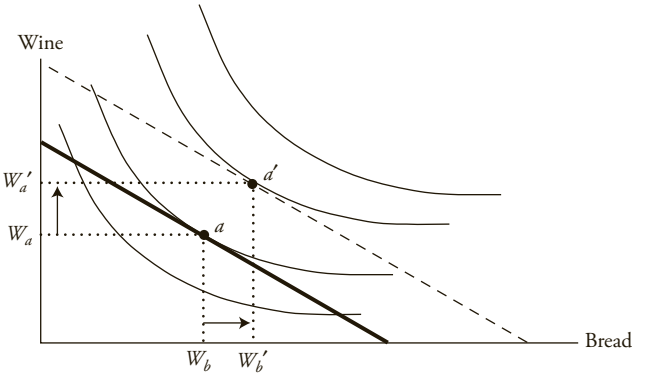
Solution to 2: The MRS_{GS} is the rate at which the consumer is willing to give up sorbet to gain a small additional amount of gelato, which is 0.75 scoops of sorbet to gain one scoop of gelato. The price ratio, P_G/P_S (0.8), is the rate at which he *must* give up sorbet to gain an additional small amount of gelato. In this case, the consumer would be better off spending a little less on gelato and a little more on sorbet.

5.2. Consumer Response to Changes in Income: Normal and Inferior Goods

Consumers' behavior is constrained by their income and the prices they must pay for the goods they consume. Consequently, if one or more of those parameters changes, consumers are likely to change their consumption behavior. We first consider an increase in income. Recall from Exhibit 2-7, Panel C, that an increase in income simply shifts the budget constraint outward from the origin, parallel to itself. Exhibit 2-12 indicates such a shift and shows how the consumer would respond, in this case, by buying more of both bread and wine.

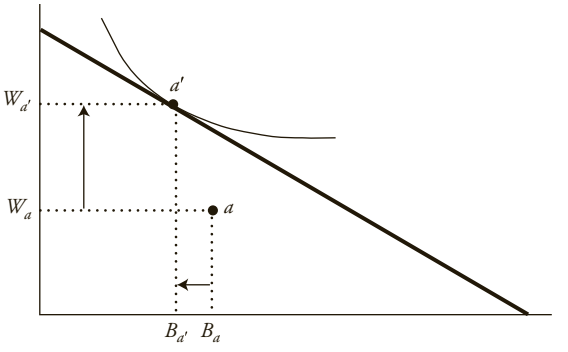
As we discovered, there is no restriction that the purchase of every good must respond to an increase in income with an increase in quantity. There, we defined *normal goods* as those with a positive response to an increase in income and *inferior goods* as those with a negative response to an increase in income. Suppose that bread is an inferior good for a particular consumer, whereas wine is a normal good. Exhibit 2-13 shows this consumer's purchase behavior when income increases. As income rises, the consumer purchases less bread but more wine.

EXHIBIT 2-12 The Effect of an Increase in Income on a Normal Good



Note: The effect of an increase in income when both goods are normal is to increase the consumption of both.

EXHIBIT 2-13 The Effect of an Increase in Income on an Inferior Good



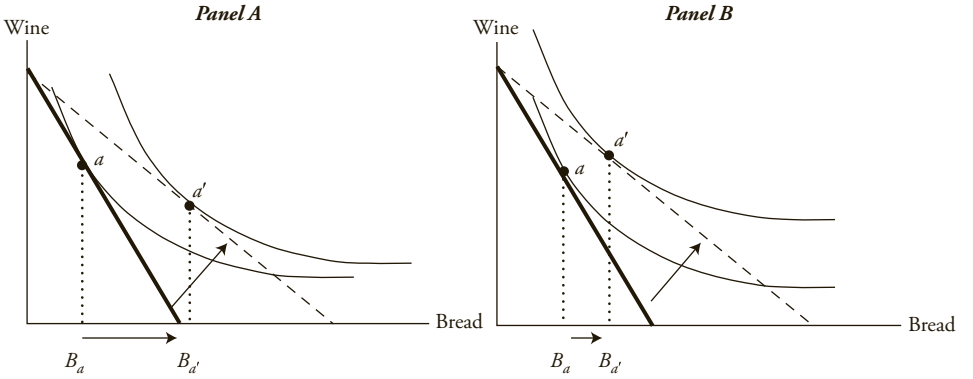
Note: The effect of an increase in income on the purchase of bread, an inferior good in this example, reduces the consumption of that good.

5.3. How the Consumer Responds to Changes in Price

We now hold income and the price of one good (wine) constant but decrease the price of the other good (bread). Recall that a decrease in the price of bread pivots the budget constraint outward along the horizontal axis but leaves the vertical intercept unchanged—as in Exhibit 2-14, where we examine two responses to the decrease in the price of bread.

In both cases, when the price of bread falls, the consumer buys more bread. But the first consumer is quite responsive to the price change, responding with an elastic demand for bread. The second consumer is still responsive, but much less so than the first consumer; this consumer's response to the price change is inelastic.

EXHIBIT 2-14 Elastic and Inelastic Responses to a Decrease in Price



Note: When the price of bread falls, this consumer moves from bundle *a* to bundle *a′*, showing an elastic response.

Note: When the price of bread falls, this consumer moves from bundle *a* to bundle *a′*, showing an inelastic response.

6. REVISITING THE CONSUMER'S DEMAND FUNCTION

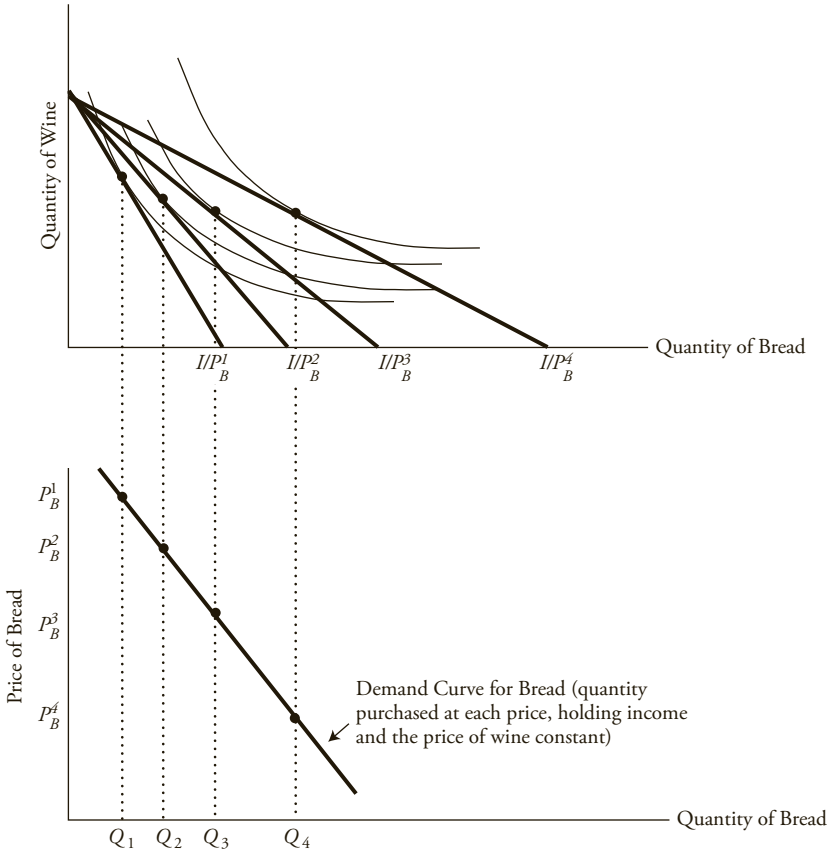
We have now come to the reason why we wanted to explore consumer theory in the first place: We want to have a sound theoretical foundation for our use of consumer demand curves. Although we could merely assume that a consumer has a demand curve, we derive a richer understanding of that curve if we start with a more fundamental recognition of the consumer's preferences and her response to changes in the parameters that constrain her behavior in the marketplace.

6.1. Consumer's Demand Curve from Preferences and Budget Constraints

Recall that to draw a consumer's demand curve, we appealed to the assumption of "holding all other things constant" and held preferences, income, and the prices of all but one good constant. Graphically, we show such an exercise by representing a given utility function with a set of indifference curves, and then we superimpose a set of budget constraints, each one representing a different price of one of the goods. Exhibit 2-15 shows the result of this exercise. Notice that we are stacking two charts vertically to show both the indifference curves and budget constraints and the demand curve below them. In the upper chart, we have rotated the budget constraint rightward, indicating successively lower prices of bread, P_B^1, P_B^2, P_B^3, P_B^4, while holding income constant at I.

This pair of diagrams deserves careful inspection. Notice first that the vertical axes are not the same. In the upper diagram, we represent the quantity of the *other* good, wine, whose price is being held constant, along with income. Hence, the budget constraints all have the same vertical intercept. But the price of bread is falling as we observe ever less steep budget constraints with horizontal intercepts moving rightward. Confronted with each respective budget constraint, the consumer finds the tangent point as indicated. This point corresponds to the respective quantities of bread, Q_1, Q_2, Q_3, and Q_4. Note also that the horizontal axes of the two diagrams are identical; they measure the quantity of bread purchased. Importantly, the

EXHIBIT 2-15 Deriving a Demand Curve



Note: A demand curve for bread is derived from the indifference curve map and a set of budget constraints representing different prices of bread.

vertical axis in the lower diagram measures the price of bread. As the price of bread falls, this consumer chooses to buy ever greater quantities, as indicated. The price–quantity combinations that result trace out this consumer's demand curve for bread in the lower diagram. For each tangent point in the upper diagram, there is a corresponding point in the lower diagram, tracing out the demand curve for bread.

6.2. Substitution and Income Effects for a Normal Good

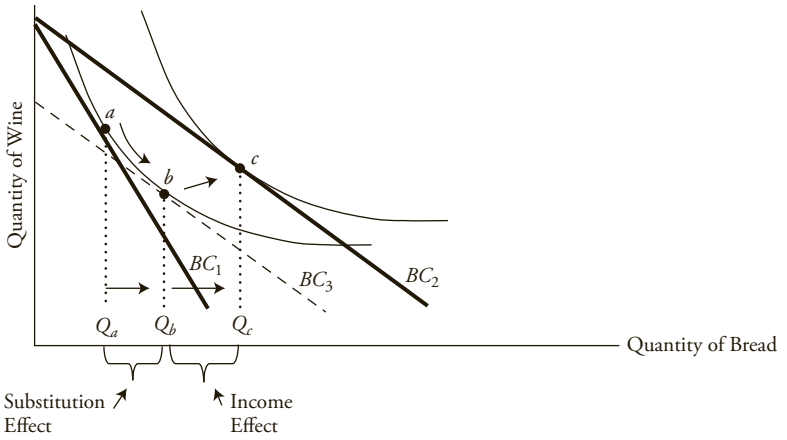
The law of demand says that when price falls, quantity demanded rises; however, it doesn't say why. We can answer that question by delving a little more deeply into consumer theory. When there is a decline in the price of a good that the consumer has been buying, two things happen. The good now becomes less costly relative to other goods. That is, it becomes more of a bargain than other things the consumer could purchase; thus, more of this good gets substituted for other goods in the consumer's market basket. Additionally, though, with the

decline in that price, the consumer's **real income** rises. We're not saying that the size of the consumer's paycheck changes; we're saying that the amount of goods that can be purchased with the same amount of money has increased. If this good is a normal good, then increases in income lead to increased purchases of this good. So the consumer tends to buy more when price falls for both reasons: the substitution effect and the income effect of a change in the price of a good.

A close look at indifference curves and budget constraints can demonstrate how these effects can be separated. Consider Exhibit 2-16, where we analyze Warren's response to a decrease in the price of bread. When the price of bread falls, as indicated by the pivoting in budget constraints from BC_1 to BC_2, Warren buys more bread, increasing his quantity from Q_a to Q_c. That is the net effect of both the substitution effect and the income effect. We can see the partial impact of each of these effects by engaging in a mental exercise. Part of Warren's response is because of his increase in real income. We can remove that effect by subtracting some income from him, while leaving the new lower price in place. The dashed budget constraint shows the reduction in income that would be just sufficient to move Warren back to his original indifference curve. Notice that we are moving BC_2 inward, parallel to itself until it becomes just tangent to Warren's original indifference curve at point *b*. The price decrease was a good thing for Warren. An offsetting bad thing would be an income reduction. If the income reduction is just sufficient to leave Warren as well off but no better off than before the price change, then we have effectively removed the real income effect of the decrease in price. What's left of his response must be due to the pure substitution effect alone. So, we say that the substitution effect is shown by the move from point *a* to point *b*. If his income reduction were then restored, the resulting movement from point *b* to point *c* must be the pure income effect.

An important thing to notice is that the pure substitution effect must always be in the direction of purchasing more when the price falls and purchasing less when the price rises. This is because of the diminishing marginal rate of substitution, or the convexity of the indifference curve. Look again at Exhibit 2-16. Note that the substitution effect is the result of changing from budget constraint 1 to budget constraint 3—or moving the budget constraint

EXHIBIT 2-16 Substitution and Income Effects for a Normal Good



Note: Substitution effect (Q_a to Q_b) and the income effect (Q_b to Q_c) of a decrease in the price of a normal good.

along the original indifference curve while maintaining tangency. Note that in the process, the budget constraint becomes less steep, just as the marginal rate of substitution decreases. Warren is no better off than before the changes, but his behavior has changed: He now buys more bread and less wine than before the offsetting changes in income and price. This reason for negatively sloped demand curves never changes.

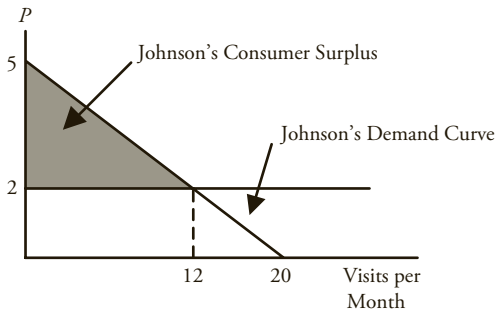
Sellers can sometimes use income and substitution effects to their advantage. Think of something you often buy, perhaps lunch at your favorite café. How much would you be willing to pay for a lunch club membership card that would allow you to purchase lunches at, say, half price? If the café could extract from you the maximum amount each month that you would be willing to pay for the half-price option, then it would successfully have removed the income effect from you in the form of a monthly fixed fee. Notice that Exhibit 2-16 implies that you would end up buying more lunches each month than before you purchased the discount card, even though you would be no better or worse off than before. This is a way that sellers are sometimes able to extract consumer surplus by means of creative pricing schemes. It's a common practice among big box retailers, sports clubs, and other users of what is called "two-part tariff pricing."

EXAMPLE 2-6 Two-Part Tariff Pricing

Nicole Johnson's monthly demand for visits to her health club is given by the following equation: $Q^d = 20 - 4P$, where Q^d is visits per month and P is euros per visit. The health club's marginal cost is fixed at €2 per visit.

1. Determine Johnson's demand curve for health club visits per month.
2. Calculate how many visits Johnson would make per month if the club charged a price per visit equal to its marginal cost.
3. Calculate Johnson's consumer surplus at the price determined in Question 2.
4. Calculate how much the club could charge Johnson each month for a membership fee.

Solution to 1: $Q^d = 20 - 4P$, so when $P = 0$, $Q^d = 20$. Inverting, $P = 5 - 0.25Q$, so when $Q = 0$, $P = 5$.



Solution to 2: $Q^d = 20 - 4(2) = 12$. Johnson would make 12 visits per month at a price of €2 per visit.

Solution to 3: Johnson's consumer surplus can be measured as the area under her demand curve and above the price she pays, for a total of 12 visits: $CS = (½)(12)(3) = 18$. Johnson would enjoy €18 per month consumer surplus.

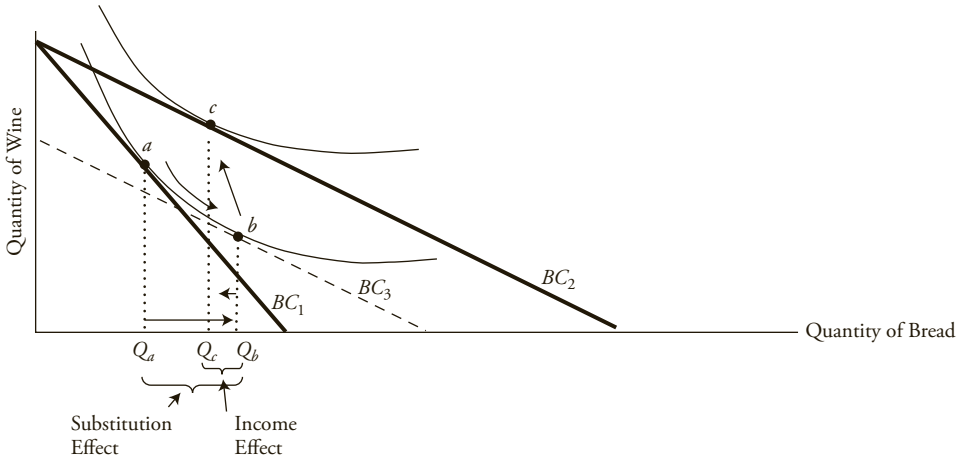
Solution to 4: The club could extract all of Johnson's consumer surplus by charging her a monthly membership fee of €18 plus a per-visit price of €2. This is called a two-part tariff because it assesses one price per unit of the item purchased plus a per-month fee (sometimes called an entry fee) equal to the buyer's consumer surplus evaluated at the per-unit price.

6.3. Income and Substitution Effects for an Inferior Good

We know that for some consumers and some goods, an increase in income leads to a decrease in the quantity purchased at each price. These goods are called *inferior* goods, and they have negative income elasticity of demand. When price falls, these goods still exhibit substitution and income effects, but they are in opposite directions. Consider Exhibit 2-17, in which we see a fairly standard set of indifference curves and budget constraints. But in this case, bread is an inferior good.

Notice that when the bread's price falls, as indicated by the shift from budget constraint 1 (BC_1) to budget constraint 2 (BC_2), the consumer buys more bread, just as we would expect. That is, the consumer's demand curve is still negatively sloped. When we apply the income adjustment to isolate substitution effect from income effect, we shift the budget constraint back to budget constraint 3, reducing income sufficiently to place the consumer back on the original indifference curve. As before, the substitution effect is shown as a movement along

EXHIBIT 2-17 Income and Substitution Effects for an Inferior Good



Note: The income effect of a price decrease for an inferior good is opposite to the substitution effect, tending to mitigate the change in quantity.

the original indifference curve from point *a* to point *b*. The income effect is, as before, a movement from one indifference curve to the other, as shown by the movement from point *b* to point *c*. In this case, however, the income effect partially offsets the substitution effect, causing demand to be less elastic than if the two effects reinforced each other.

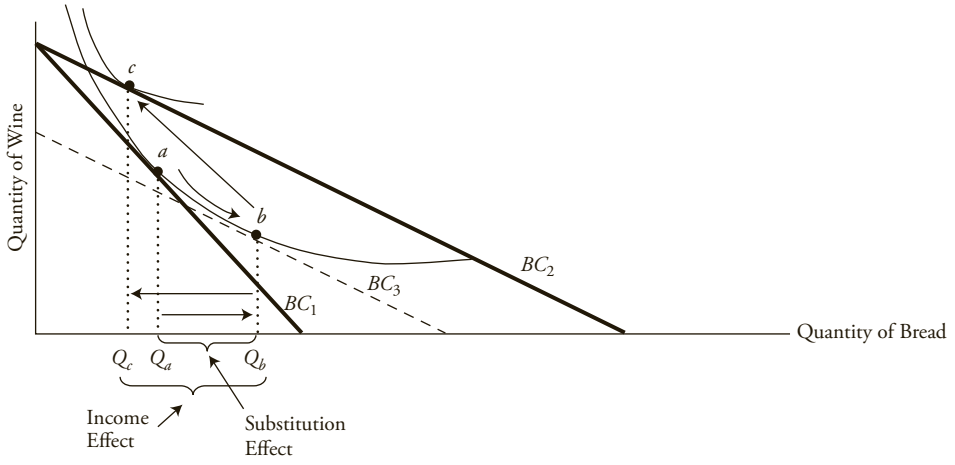
We see that for inferior goods, the income effect and the substitution effect are in opposite directions: The decrease in price causes the consumer to buy more, but the income effect tends to mitigate that effect. It's still true that a decrease in the price of bread represents an increase in real income. But in the case of an inferior good, the increased income causes the consumer to want to buy less of the good, not more. As long as the income effect has a lower magnitude than the substitution effect, the consumer still ends up buying more at the lower price. However, the consumer buys a little less than she would if the good were normal. It is possible, though highly unlikely, for the income effect to have greater magnitude than the substitution effect. We examine that case next.

In the case of savings, the same type of effects can apply. For example, say interest rates rise. Individuals may save more because the reward (price) for saving has risen, and individuals substitute future consumption for present consumption. However, higher interest rates also imply that less saving is required to attain a given future amount of money. If the latter effect (the income effect) dominates, then it is possible to observe higher interest rates resulting in less savings.

6.4. Negative Income Effect Larger than Substitution Effect: Giffen Goods

In theory, it is possible for the income effect to be so strong and so negative as to overpower the substitution effect. If that were to occur, then a decrease in price could result in a *decrease* in quantity demanded and a *positively sloped* individual demand curve. Let us explore this curiosity in Exhibit 2-18.

EXHIBIT 2-18 Income and Substitution Effects for a Giffen Good



Note: Income and substitution effects of a fall in price for a Giffen good: When price declines, the consumer chooses to buy less of the good.

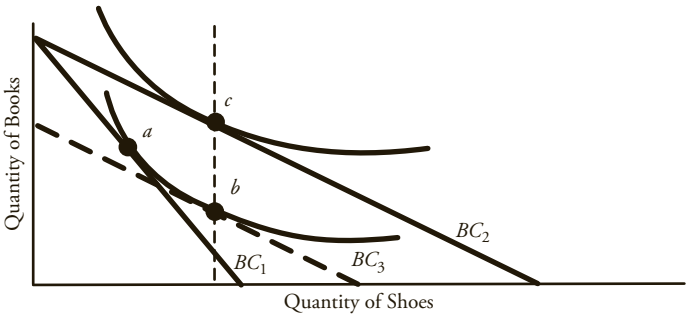
Once again, we decrease the price of bread as indicated by the pivoting of the budget constraint form BC_1 to BC_2, and then we move the budget constraint parallel to itself leftward until it just touches the original indifference curve at point *b* to remove the income effect. What is left is the substitution effect. As always, the substitution effect causes the consumer to substitute more bread for less wine in the basket, as indicated by the movement along the indifference curve from point *a* to point *b*. But notice the odd result when we "give back" the income and move from BC_3 to BC_2. The income effect for this inferior good (from point *b* to point *c*) is once again opposite in direction to the substitution effect, as is true for all inferior goods. But in this curious case, its magnitude overwhelms the substitution effect: point *c* lies *to the left* of point *a*. The consumer actually buys less of the good when its price falls, resulting in a positively sloped demand curve. If we reversed the analysis and increased the price of bread, this consumer would buy more bread when its price rose. Those inferior goods whose income effect is negative and greater in magnitude than the substitution effect are known as **Giffen goods**. Importantly, all Giffen goods must be inferior, but not all inferior goods are Giffen goods.

This curious result is originally attributed to Robert Giffen, who suspected that Irish peasants might have responded this way to increased prices of staples during the Irish potato famine in the nineteenth century. He reasoned that staples such as potatoes comprised a very large portion of the peasants' total budgets. Additionally, potatoes could be a very inferior good, which simply means that when incomes fell, the peasants bought a lot more potatoes; and when incomes rose, they bought a lot fewer. Now, because potatoes took up such a large part of total expenditures, any increase in the price of potatoes would result in a very substantial decrease in real income. This combination of strong inferiority coupled with a large amount of the budget spent on potatoes could, in theory, result in the negative income effect not only being opposite in direction to the substitution effect, but in fact overwhelming it.

Although some empirical studies have suggested the existence of Giffen goods, even if they existed they would be extraordinarily rare. Moreover, although they might exist for some small subset of consumers, it is highly unlikely that consumers as a whole would behave this way. So Giffen goods' role in microeconomics is greater than their role in the empirical world. True, they result in a positively sloped demand curve and they do not violate any of the axioms of consumer choice theory. But any company managers who believe that if they raise the price of their product they will sell *more* of it are very likely to be disappointed.

EXAMPLE 2-7 Income and Substitution Effects of a Decrease in Price

Consider the following diagram of budget constraints and indifference curves for a consumer choosing to allocate her budget between books and shoes. Determine whether shoes are normal, inferior, or Giffen goods for this consumer.



Solution: When the price of shoes falls, the original budget constraint pivots from BC_1 to BC_2. The original tangent point was at *a* and is now at *c*. We can separate the substitution effect from the income effect by removing enough income to put the consumer back on the old indifference curve at point *b*. This is shown as a shift in the budget constraint from BC_2 to BC_3, a parallel shift. The substitution effect is, therefore, from point *a* to point *b*, along the original indifference curve. The income effect is from point *b* to point *c*, but note that those two points are on the same vertical line. In this case, shoes are on a borderline between normal and inferior. There is zero income effect. (If point *c* had been to the right of point *b*, shoes would be normal. And if point *c* had been to the left of point *b*, they would be inferior. Finally, if point *c* had been to the left of point *a*, shoes would be a Giffen good.)

6.5. Veblen Goods: Another Possibility for a Positively Sloped Demand Curve

Standard choice theory assumes that the consumer can always make comparisons among all pairs of bundles of goods and identify preferences before knowing anything about the prices of those baskets. Then, as we've seen, the consumer is constrained by income and prices and makes actual choices of which bundles to purchase with limited income. It is important to note that those preferences are assumed to exist even before knowing the prices at which those goods could be purchased. It is possible, however, that an item's price tag *itself* might help determine the consumer's preferences for it. Thorstein Veblen posited just such a circumstance in his concept of **conspicuous consumption**. According to this way of thinking, a consumer might derive utility out of being known by others to consume a so-called *high-status* good, such as a luxury automobile or a very expensive piece of jewelry. Importantly, it is the high price itself that partly imparts value to such a good. If that is the case, then a consumer would actually value a good more if it had a higher price. So, it is argued that an increase in the price of a **Veblen good** would result in the consumer being *more inclined* to purchase it, not less. In the extreme, it could be argued that the consumer's demand for such a good could be positively sloped, though this need not necessarily follow. In fact, of course, if any seller actually faced a positively sloped demand curve for a product, the rational response would be to increase price because more would be sold at the higher price. Ultimately, at some very high price, demand would necessarily become negatively sloped.

It is important to recognize that, although Veblen goods and Giffen goods share some characteristics, they are in fact quite different. Whether or not they actually exist, Giffen goods are not inconsistent with the fundamental axioms of demand theory. True, they would result in a violation of the law of demand, but that law is not a logical necessity. It is simply recognition that in virtually all observed cases, demand curves are negatively sloped. Giffen goods certainly would not be considered examples of status goods, because an increase in income alone would result in a reduced interest in purchasing them. Veblen goods, in contrast, derive their value from the ostentatious consumption of them as symbols of the purchaser's high status in society. If they exist, they are certainly not inferior goods. And they do violate the axioms of choice that form the foundation of accepted demand theory.

7. SUMMARY

This chapter has explored how consumer preferences over baskets of goods and budget constraints translate into the demand curves posited by the demand and supply model of markets. Among the major points made are the following:

- Consumer choice theory is the branch of microeconomics that relates consumer demand curves to consumer preferences. Utility theory is a quantitative model of consumer preferences and is based on a set of axioms (assumptions that are assumed to be true). If consumer preferences are complete, transitive, and insatiable, those preferences can be represented by an ordinal utility function and depicted by a set of indifference curves that are generally negatively sloped, are convex from below, and do not cross for a given consumer.
- A consumer's relative strength of preferences can be inferred from his marginal rate of substitution of good X for good Y (MRS_{XY}), which is the rate at which the consumer is willing to sacrifice good Y to obtain an additional small increment of good X. If two consumers have different marginal rates of substitution, they can both benefit from the voluntary exchange of one good for the other.
- A consumer's attainable consumption options are determined by her income and the prices of the goods she must purchase to consume. The set of options available is bounded by the budget constraint, a negatively sloped linear relationship that shows the highest quantity of one good that can be purchased for any given amount of the other good being bought.
- Analogous to the consumer's consumption opportunity set are, respectively, the production opportunity set and the investment opportunity set. A company's production opportunity set represents the greatest quantity of one product that a company can produce for any given amount of the other good it produces. The investment opportunity set represents the highest return an investor can expect for any given amount of risk undertaken.
- Consumer equilibrium is obtained when utility is maximized, subject to the budget constraint, generally depicted as a tangency between the highest attainable indifference curve and the fixed budget constraint. At that tangency, the MRS_{XY} is just equal to the two goods' price ratio, P_X/P_Y—or that bundle such that the rate at which the consumer is just willing to sacrifice good Y for good X is equal to the rate at which, based on prices, she must sacrifice good Y for good X.
- If the consumer's income and the price of all other goods are held constant and the price of good X is varied, the set of consumer equilibria that results will yield that consumer's demand curve for good X. In general, we expect the demand curve to have a negative slope (the law of demand) because of two influences: income and substitution effects of a decrease

in price. Normal goods have a negatively sloped demand curve. For normal goods, income and substitution effects reinforce one another. However, for inferior goods, the income effect offsets part or all of the substitution effect. In the case of the Giffen good, the income effect of this very inferior good overwhelms the substitution effect, resulting in a positively sloped demand curve.

- In accepted microeconomic consumer theory, the consumer is assumed to be able to judge the value of any given bundle of goods without knowing anything about their prices. Then, constrained by income and prices, the consumer is assumed to be able to choose the optimal bundle of goods that is in the set of available options. It is possible to conceive of a situation in which the consumer cannot truly value a good until the price is known. In these Veblen goods, the price is used by the consumer to signal the consumer's status in society. Thus, to some extent, the higher the price of the good, the more value it offers to the consumer. In the extreme case, this could possibly result in a positively sloped demand curve. This result is similar to a Giffen good, but the two goods are fundamentally different.

PRACTICE PROBLEMS[1]

1. A child indicates that she prefers going to the zoo over the park and prefers going to the beach over the zoo. When given the choice between the park and the beach, she chooses the park. Which of the following assumptions of consumer preference theory is she *most likely* violating?
   A. Nonsatiation
   B. Complete preferences
   C. Transitive preferences

2. Which of the following ranking systems *best* describes consumer preferences within a utility function?
   A. Util
   B. Ordinal
   C. Cardinal

3. Which of the following statements *best* explains why indifference curves are generally convex as viewed from the origin?
   A. The assumption of nonsatiation results in convex indifference curves.
   B. The marginal rate of substitution of one good for another remains constant along an indifference curve.
   C. The marginal utility gained from one additional unit of a good versus another diminishes the more one has of the first good.

4. If a consumer's marginal rate of substitution of good X for good Y (MRS_{XY}) is equal to 2, then the:
   A. consumer is willing to give up two units of X for one unit of Y.
   B. slope of a line tangent to the indifference curve at that point is 2.
   C. slope of a line tangent to the indifference curve at that point is -2.

[1]These practice problems were written by William Akmentins, CFA (Dallas, Texas, USA).

5. In the case of two goods, *x* and *y*, which of the following statements is *most likely* true? Maximum utility is achieved:
   A. along the highest indifference curve below the budget constraint line.
   B. at the tangency between the highest attainable indifference curve and the budget constraint line.
   C. when the marginal rate of substitution is equal to the ratio of the price of good *y* to the price of good *x*.

6. In the case of a normal good with a decrease in its own price, which of the following statements is *most likely* true?
   A. Both the substitution effect and the income effect lead to an increase in the quantity purchased.
   B. The substitution effect leads to an increase in the quantity purchased, while the income effect has no impact.
   C. The substitution effect leads to an increase in the quantity purchased, while the income effect leads to a decrease.

7. For a Giffen good, the:
   A. demand curve is positively sloped.
   B. substitution effect overwhelms the income effect.
   C. income and substitution effects are in the same direction.

8. Which of the following statements *best* illustrates the difference between a Giffen good and a Veblen good?
   A. The Giffen good alone is an inferior good.
   B. Their substitution effects are in opposite directions.
   C. The Veblen good alone has a positively sloped demand curve.

CHAPTER 3

DEMAND AND SUPPLY ANALYSIS: THE FIRM

Gary L. Arbogast, CFA

Richard V. Eastin

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Calculate, interpret, and compare accounting profit, economic profit, normal profit, and economic rent.
- Calculate, interpret, and compare total, average, and marginal revenue.
- Describe the firm's factors of production.
- Calculate and interpret total, average, marginal, fixed, and variable costs.
- Determine and describe breakeven and shutdown points of production.
- Explain how economies of scale and diseconomies of scale affect costs.
- Describe approaches to determining the profit-maximizing level of output.
- Distinguish between short-run and long-run profit maximization.
- Distinguish among decreasing-cost, constant-cost, and increasing-cost industries and describe the long-run supply of each.
- Calculate and interpret total, marginal, and average product of labor.
- Describe the phenomenon of diminishing marginal returns, and calculate and interpret the profit-maximizing utilization level of an input.
- Determine the optimal combination of resources that minimizes cost.

1. INTRODUCTION

In studying decision making by consumers and businesses, microeconomics gives rise to the theory of the consumer and the theory of the firm as two branches of study.

The **theory of the consumer** is the study of consumption—the demand for goods and services—by utility-maximizing individuals. The **theory of the firm**, the subject of this

chapter, is the study of the supply of goods and services by profit-maximizing firms. Conceptually, profit is the difference between revenue and costs. Revenue is a function of selling price and quantity sold, which are determined by the demand and supply behavior in the markets into which the firm sells or provides its goods or services. Costs are a function of the demand and supply interactions in resource markets, such as markets for labor and for physical inputs. The main focus of this chapter is the cost side of the profit equation for companies competing in market economies under perfect competition. The next chapter examines the different types of markets into which a firm may sell its output.

The study of the profit-maximizing firm in a single time period is the essential starting point for the analysis of the economics of corporate decision making. Furthermore, with the attention given to earnings by market participants, the insights gained by this study should be practically relevant. Among the questions this chapter will address are the following:

- How should profit be defined from the perspective of suppliers of capital to the firm?
- What is meant by factors of production?
- How are total, average, and marginal costs distinguished, and how is each related to the firm's profit?
- What roles do marginal quantities (selling prices and costs) play in optimization?

This chapter is organized as follows: Section 2 discusses the types of profit measures, including what they have in common, how they differ, and their uses and definitions. Section 3 covers the revenue and cost inputs of the profit equation and the related topics of breakeven analysis, shutdown point of operation, market entry and exit, cost structures, and scale effects. In addition, the economic outcomes related to a firm's optimal supply behavior over the short run and the long run are presented in this section. A summary and practice problems conclude the chapter.

2. OBJECTIVES OF THE FIRM

This chapter assumes that the objective of the firm is to maximize profit over the period ahead. Such analysis provides both tools (e.g., optimization) and concepts (e.g., productivity) that can be adapted to more complex cases, and also provides a set of results that may offer useful approximations in practice. The price at which a given quantity of a good can be bought or sold is assumed to be known with certainty (i.e., the theory of the firm under conditions of certainty). The main contrast of this type of analysis is to the theory of the firm under conditions of uncertainty, where prices, and therefore profit, are uncertain. Under market uncertainty, a range of possible profit outcomes is associated with the firm's decision to produce a given quantity of goods or services over a specific time period. Such complex theory typically makes simplifying assumptions. When managers of for-profit companies have been surveyed about the objectives of the companies they direct, researchers have often concluded that (1) companies frequently have multiple objectives; (2) objectives can often be classified as focused on profitability (e.g., maximizing profits, increasing market share) or on controlling risk (e.g., survival, stable earnings growth); and (3) managers in different countries may have different emphases.

Finance experts frequently reconcile profitability and risk objectives by stating that the objective of the firm is, or should be, **shareholder wealth maximization** (i.e., to maximize the market value of shareholders' equity). This theory states that firms try, or should try, to

increase the wealth of their owners (shareholders) and that market prices balance returns against risk. However, complex corporate objectives may exist in practice. Many analysts view profitability as the single most important measure of business performance. Without profit, the business eventually fails; with profit, the business can survive, compete, and prosper. The question is: What is profit? Economists, accountants, investors, financial analysts, and regulators view profit from different perspectives. The starting point for anyone who is doing profit analysis is to have a solid grasp of how various forms of profit are defined and how to interpret the profit based on these different definitions.

By defining profit in general terms as the difference between total revenue and total costs, profit maximization involves the following expression:

$$\Pi = TR - TC \tag{3-1}$$

where Π is profit, TR is total revenue, and TC is total costs. TC can be defined as accounting costs or economic costs, depending on the objectives and requirements of the analyst for evaluating profit. The characteristics of the product market, where the firm sells its output or services, and of the resource market, where the firm purchases resources, play an important role in the determination of profit. Key variables that determine TC are the level of output, the firm's efficiency in producing that level of output when utilizing inputs, and resource prices as established by resource markets. TR is a function of output and product price as determined by the firm's product market.

2.1. Types of Profit Measures

The economics discipline has its own concept of profit, which differs substantially from what accountants consider profit. There are thus two basic types of profit—accounting and economic—and analysts need to be able to interpret each correctly and to understand how they are related to each other. In the theory of the firm, however, *profit* without further qualification refers to *economic profit*.

2.1.1. Accounting Profit

Accounting profit is generally defined as net income reported on the income statement according to standards established by private and public financial oversight bodies that determine the rules for calculating accounting profit. One widely accepted definition of accounting profit—also known as net profit, net income, or net earnings—states that it equals revenue less all **accounting (or explicit) costs**. Accounting or explicit costs are payments to nonowner parties for services or resources that they supply to the firm. Often referred to as the "bottom line" (the last income figure in the income statement), accounting profit is what is left after paying all accounting costs—regardless of whether the expense is a cash outlay. When accounting profit is negative, it is called an **accounting loss**. Equation 3-2 summarizes the concept of accounting profit:

$$\text{Accounting profit} = \text{Total revenue} - \text{Total accounting costs} \tag{3-2}$$

When defining profit as accounting profit, the TC term in Equation 3-1 becomes total accounting costs, which include only the explicit costs of doing business. Let us consider two businesses: a start-up company and a publicly traded corporation. Suppose that for the start-up, total revenue in the business's first year is €3,500,000 and total accounting costs are

€3,200,000. Accounting profit is €3,500,000 − €3,200,000 = €300,000. The corresponding calculation for the publicly traded corporation, let us suppose, is $50,000,000 − $48,000,000 = $2,000,000. Note that total accounting costs in either case include interest expense—which represents the return required by suppliers of debt capital—because interest expense is an explicit cost.

2.1.2. Economic Profit and Normal Profit

Economic profit (also known as **abnormal profit** or **supernormal profit**) may be defined broadly as accounting profit less the implicit opportunity costs not included in total accounting costs:

$$\text{Economic profit} = \text{Accounting profit} - \text{Total implicit opportunity costs} \quad (3\text{-}3a)$$

We can define a term, **economic cost**, equal to the sum of total accounting costs and implicit opportunity costs. Economic profit is therefore equivalently defined as:

$$\text{Economic profit} = \text{Total revenue} - \text{Total economic costs} \quad (3\text{-}3b)$$

For publicly traded corporations, the focus of investment analysts' work, the cost of equity capital is the largest and most readily identified implicit opportunity cost omitted in calculating total accounting cost. Consequently, economic profit can be defined for publicly traded corporations as accounting profit less the required return on equity capital.

Examples will make these concepts clearer. Consider the start-up company for which we calculated an accounting profit of €300,000 and suppose that the entrepreneurial executive who launched the start-up took a salary reduction of €100,000 per year relative to the job he left. That €100,000 is an opportunity cost of involving him in running the start-up. Besides labor, financial capital is a resource. Suppose that the executive, as sole owner, makes an investment of €1,500,000 to launch the enterprise and that he might otherwise expect to earn €200,000 per year on that amount in an investment with similar risk. Total implicit opportunity costs are €100,000 + €200,000 = €300,000 per year and economic profit is zero: €300,000 − €300,000 = €0. For the publicly traded corporation, we consider the cost of equity capital as the only implicit opportunity cost identifiable. Suppose that equity investment is $18,750,000 and shareholders' required rate of return is 8 percent so that the dollar cost of equity capital is $1,500,000. Economic profit for the publicly traded corporation is therefore $2,000,000 (accounting profit) less $1,500,000 (cost of equity capital) or $500,000.

For the start-up company, economic profit was zero. Total economic costs were just covered by revenues, and the company was not earning a euro more or less than the amount that met the opportunity costs of the resources used in the business. Economists would say the company was earning a normal profit (economic profit of zero). In simple terms, **normal profit** is the level of accounting profit needed to just cover the implicit opportunity costs ignored in accounting costs. For the publicly traded corporation, normal profit was $1,500,000: normal profit can be taken to be the cost of equity capital (in money terms) for such a company or the dollar return required on an equal investment by equity holders in an equivalently risky alternative investment opportunity. The publicly traded corporation actually earned $500,000 in excess of normal profit, which should be reflected in the common shares' market price.

Thus, the following expression links accounting profit to economic profit and normal profit:

$$\text{Accounting profit} = \text{Economic profit} + \text{Normal profit} \tag{3-4}$$

When accounting profit equals normal profit, economic profit is zero. Further, when accounting profit is greater than normal profit, economic profit is positive; and when accounting profit is less than normal profit, economic profit is negative (the firm has an **economic loss**).

Economic profit for a firm can originate from sources such as:

- Competitive advantage.
- Exceptional managerial efficiency or skill.
- Difficult-to-copy technology or innovation (e.g., patents, trademarks, and copyrights).
- Exclusive access to less expensive inputs.
- Fixed supply of an output, commodity, or resource.
- Preferential treatment under governmental policy.
- Large increases in demand where supply is unable to respond fully over time.
- Exertion of monopoly power (price control) in the market.
- Market barriers to entry that limit competition.

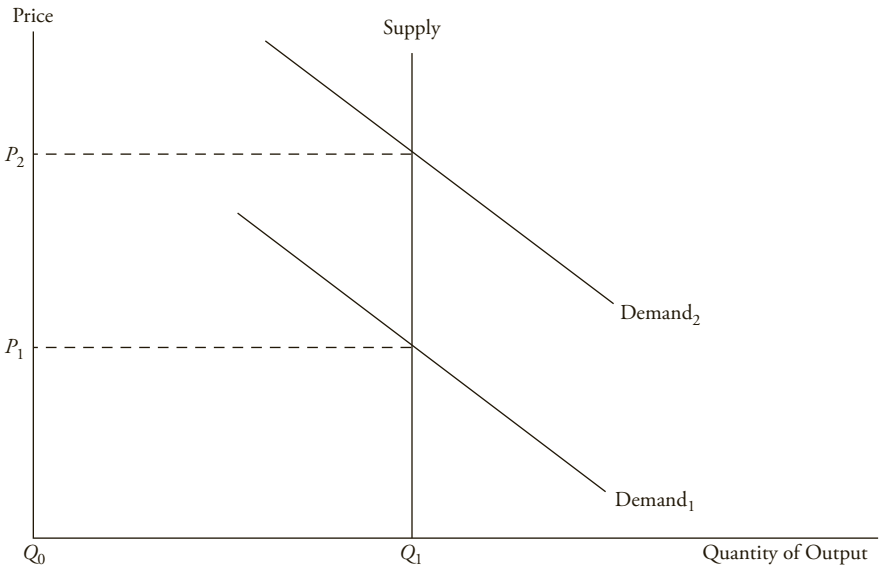
Any of these factors may lead the firm to have positive net present value (NPV) investment opportunities. Access to positive NPV opportunities and therefore profit in excess of normal profits in the short run may or may not exist in the long run, depending on the potential strength of competition. In highly competitive market situations, firms tend to earn the normal profit level over time because ease of market entry allows for other competing firms to compete away any economic profit over the long run. Economic profit that exists over the long run is usually found where competitive conditions persistently are less than perfect in the market.

2.1.3. Economic Rent

The surplus value known as **economic rent** results when a particular resource or good is fixed in supply (with a vertical supply curve) and market price is higher than what is required to bring the resource or good onto the market and sustain its use. Essentially, demand determines the price level and the magnitude of economic rent that is forthcoming from the market. Exhibit 3-1 illustrates this concept, where P_1 is the price level that yields a normal profit return to the business that supplies the item. When demand increases from Demand_1 to Demand_2, price rises to P_2, where at this higher price level economic rent is created. The amount of this economic rent is calculated as $(P_2 - P_1) \times Q_1$. The firm has not done anything internally to merit this special reward: It benefits from an increase in demand in conjunction with a supply curve that does not fully adjust with an increase in quantity when the price rises.

Because of their limited availability in nature, certain resources—such as land and specialty commodities—possess highly **inelastic supply** curves in both the short run and the long run (shown in Exhibit 3-1 as a vertical supply curve). When supply is relatively inelastic, a high degree of market demand can result in pricing that creates economic rent. This economic rent results from the fact that when price increases, the quantity supplied does not change or, at the most, increases only slightly. This is because of the fixation of supply by nature or by such artificial constraints as government policy.

EXHIBIT 3-1 Economic Rent



How is the concept of economic rent useful in financial analysis? Commodities or resources that command economic rent have the potential to reward equity investors more than what is required to attract their capital to that activity, resulting in greater shareholders' wealth. Evidence of economic rent attracts additional capital funds to the economic endeavor. This new investment capital increases shareholders' value as investors bid up share prices of existing firms. Any commodity, resource, or good that is fixed or nearly fixed in supply has the potential to yield economic rent. From an analytical perspective, one can obtain industry supply data to calculate the **elasticity of supply**, which measures the sensitivity of quantity supplied to a change in price. If quantity supplied is relatively unresponsive (**inelastic**) to price changes, then a potential condition exists in the market for economic rent. A reliable forecast of changes in demand can indicate the degree of any economic rent that is forthcoming from the market in the future. When one is analyzing fixed or nearly fixed supply markets (e.g., gold), a fundamental comprehension of demand determinants is necessary to make rational financial decisions based on potential economic rent.

EXAMPLE 3-1 Economic Rent and Investment Decision Making

The following market data show the global demand, global supply, and price on an annual basis for gold over the period 2006–2008. Based on the data, what observation can be made about market demand, supply, and economic rent?

Year	2006	2007	2008	Percent Change 2006–2008
Supply (in metric tons)	3,569	3,475	3,508	−1.7
Demand (in metric tons)	3,423	3,552	3,805	+11.2
Average spot price (in US$)	603.92	695.39	871.65	+44.3

Source: GFMS Ltd and World Gold Council.

Solution: The amount of total gold supplied to the world market over this period actually declined slightly by 1.7 percent during a period when there was a double-digit increase of 11.2 percent in demand. As a consequence, the spot price dramatically increased by 44.3 percent. Economic rent resulted from this market relationship of a relatively fixed supply of gold and a rising demand for it.

2.2. Comparison of Profit Measures

All three types of profit are interconnected because, according to Equation 3-4, accounting profit is the summation of normal and economic profit. In the short run, the normal profit rate is relatively stable, which makes accounting and economic profit the two variable terms in the profit equation. Over the longer term, all three types of profit are variable, where the normal profit rate can change according to investment returns across firms in the industry.

Normal profit is necessary to stay in business in the long run; positive economic profit is not. A business can survive indefinitely by just making the normal profit return for investors. Failing to earn normal profits over the long run has a debilitating impact on the firm's ability to access capital and to function properly as a business enterprise. Consequentially, the market value of equity and shareholders' wealth deteriorates whenever risk to achieving normal profit materializes and the firm fails to reward investors for their risk exposure and for the opportunity cost of their equity capital.

To summarize, the ultimate goal of analyzing the different types of profit is to determine how their relationships to one another influence the firm's market value of equity. Exhibit 3-2 compares accounting, normal, and economic profits in terms of how a firm's market value of equity is impacted by the relationships among the three types of profit.

EXHIBIT 3-2 Relationship of Accounting, Normal, and Economic Profit to Equity Value

Relationship between Accounting Profit and Normal Profit	Economic Profit	Firm's Market Value of Equity
Accounting profit > Normal profit	Economic profit > 0 and firm is able to protect economic profit over the long run	Positive effect
Accounting profit = Normal profit	Economic profit = 0	No effect
Accounting profit < Normal profit	Economic profit < 0 and implies economic loss	Negative effect

3. ANALYSIS OF REVENUE, COSTS, AND PROFITS

To fully comprehend the dimensions of profit maximization, one must have a detailed understanding of the revenue and cost variables that determine profit.

Revenue and cost flows are calculated in terms of total, average, and marginal. A total is the summation of all individual components. For example, total cost is the summation of all costs that are incurred by the business. Total revenue is the sum of the revenues from all the business's units. In the theory of the firm, averages and marginals are calculated with respect to the quantity produced and sold in a single period (as opposed to averaging a quantity over a number of time periods). For example, average revenue is calculated by dividing total revenue by the number of items sold. To calculate a marginal term, take the change in the total and divide by the change in the quantity number.

Exhibit 3-3 shows a summary of the terminology and formulas pertaining to profit maximization, where profit is defined as total revenue minus total economic costs. Note that the definition of profit is the economic version, which recognizes that the implicit opportunity costs of equity capital, in addition to explicit accounting costs, are economic costs. The first main category consists of terms pertaining to the revenue side of the profit equation: total revenue, average revenue, and marginal revenue. Cost terms follow with an overview of the different types of costs—total, average, and marginal.

EXHIBIT 3-3 Summary of Profit, Revenue, and Cost Terms

Term	Calculation
Profit	
(Economic) profit	Total revenue minus total economic cost; $TR - TC$
Revenue	
Total revenue (TR)	Price times quantity ($P \times Q$), or the sum of individual units sold times their respective prices; $\Sigma(P_i \times Q_i)$
Average revenue (AR)	Total revenue divided by quantity; $TR \div Q$
Marginal revenue (MR)	Change in total revenue divided by change in quantity; $\Delta TR \div \Delta Q$
Costs	
Total fixed cost (TFC)	Sum of all fixed expenses; here defined to include all opportunity costs
Total variable cost (TVC)	Sum of all variable expenses, or per unit variable cost times quantity; per-unit $VC \times Q$
Total costs (TC)	Total fixed cost plus total variable cost; $TFC + TVC$
Average fixed cost (AFC)	Total fixed cost divided by quantity; $TFC \div Q$
Average variable cost (AVC)	Total variable cost divided by quantity; $TVC \div Q$
Average total cost (ATC)	Total cost divided by quantity; $(TC \div Q)$ or $(AFC + AVC)$
Marginal cost (MC)	Change in total cost divided by change in quantity; $\Delta TC \div \Delta Q$

3.1. Profit Maximization

In free markets—and even in regulated market economies—profit maximization tends to promote economic welfare and a higher standard of living, and creates wealth for investors. Profit motivates businesses to use resources efficiently and to concentrate on activities in which they have a competitive advantage. Most economists believe that profit maximization promotes allocational efficiency—that resources flow into their highest-valued uses.

Overall, the functions of profit are as follows:

- Rewards entrepreneurs for risk taking when pursuing business ventures to satisfy consumer demand.
- Allocates resources to their most efficient use; input factors flow from sectors with economic losses to sectors with economic profit, where profit reflects goods most desired by society.
- Spurs innovation and the development of new technology.
- Stimulates business investment and economic growth.

There are three approaches to calculate the point of profit maximization. First, given that profit is the difference between total revenue and total costs, maximum profit occurs at the output level where this difference is the greatest. Second, maximum profit can also be calculated by comparing revenue and cost for each individual unit of output that is produced and sold. A business increases profit through greater sales as long as per-unit revenue exceeds per-unit cost on the next unit of output sold. Profit maximization takes place at the point where the last individual output unit breaks even. Beyond this point, total profit decreases because the per-unit cost is higher than the per-unit revenue from successive output units. A third approach compares the revenue generated by each resource unit with the cost of that unit. Profit contribution occurs when the revenue from an input unit exceeds its cost. The point of profit maximization is reached when resource units no longer contribute to profit. All three approaches yield the same profit-maximizing quantity of output. (These approaches will be explained in greater detail later.)

Because profit is the difference between revenue and cost, an understanding of profit maximization requires that we examine both of those components. Revenue comes from the demand for the firm's products, and cost comes from the acquisition and utilization of the firm's inputs in the production of those products.

3.1.1. Total, Average, and Marginal Revenue

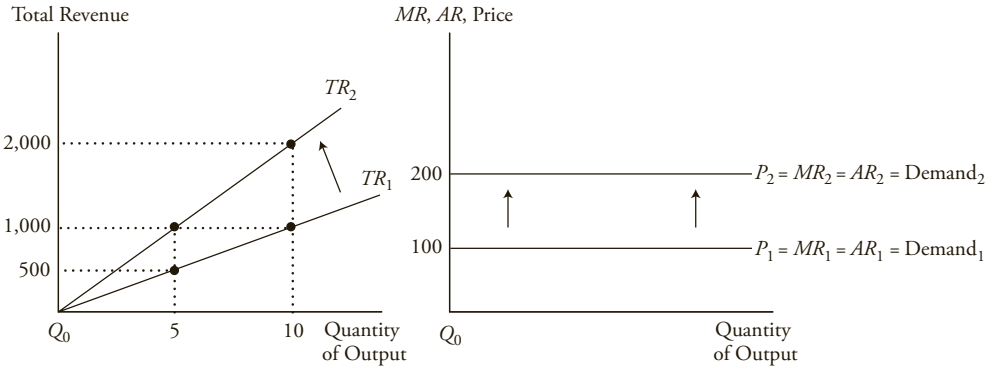
This section briefly examines demand and revenue in preparation for addressing cost. Unless the firm is a pure **monopolist** (i.e., the only seller in its market), there is a difference between market demand and the demand facing an individual firm. The next chapter devotes much more time to understanding the various competitive environments (perfect competition, monopolistic competition, oligopoly, and monopoly), known as **market structure**. To keep the analysis simple at this point, we note that competition could be either perfect or imperfect. In **perfect competition**, the individual firm has virtually no impact on market price, because it is assumed to be a very small seller among a very large number of firms selling essentially identical products. Such a firm is called a **price taker**. In the second case, the firm does have at least some control over the price at which it sells its product because it must lower its price to sell more units.

Exhibit 3-4 presents total, average, and marginal revenue data for a firm under the assumption that the firm is price taker at each relevant level of quantity of goods sold.

EXHIBIT 3-4 Total, Average, and Marginal Revenue under Perfect Competition

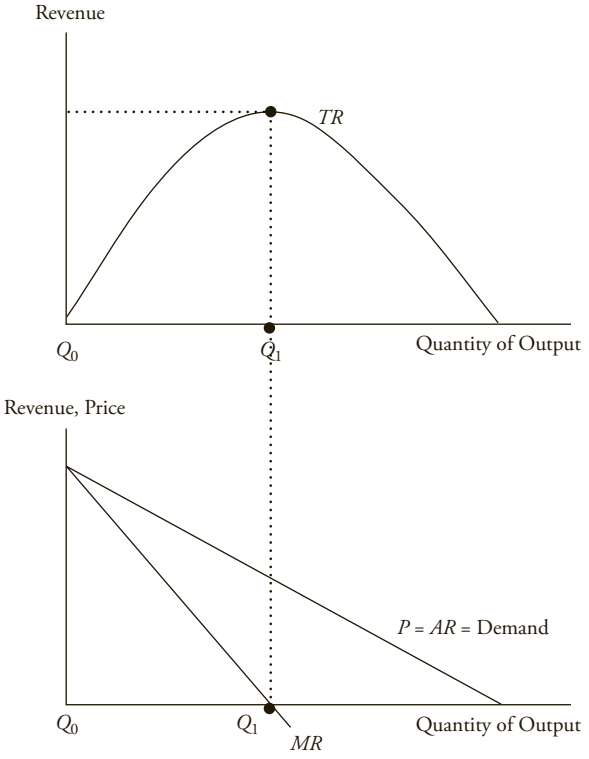
Quantity Sold (*Q*)	Price (*P*)	Total Revenue (*TR*)	Average Revenue (*AR*)	Marginal Revenue (*MR*)
0	100	0	—	—
1	100	100	100	100
2	100	200	100	100
3	100	300	100	100
4	100	400	100	100
5	100	500	100	100
6	100	600	100	100
7	100	700	100	100
8	100	800	100	100
9	100	900	100	100
10	100	1,000	100	100

EXHIBIT 3-5 Total Revenue, Average Revenue, and Marginal Revenue under Perfect Competition



Consequently, the individual seller faces a horizontal demand curve over relevant output ranges at the price level established by the market (see Exhibit 3-5). The seller can offer any quantity at this set market price without affecting price. In contrast, **imperfect competition** is where an individual firm has enough share of the market (or can control a certain segment of the market) and is therefore able to exert some influence over price. Instead of a large number of competing firms, imperfect competition involves a smaller number of firms in the market relative to perfect competition and in the extreme case only one firm (i.e., monopoly). Under any form of imperfect competition, the individual seller confronts a negatively sloped demand curve, where price and the quantity demanded by consumers are inversely related. In this case, price to the firm declines when a greater quantity is offered to the market; price to the firm increases when a lower quantity is offered to the market. This is shown in Exhibits 3-6 and 3-7.

EXHIBIT 3-6 Total Revenue, Average Revenue, and Marginal Revenue under Imperfect Competition



EXAMPLE 3-2 Calculation and Interpretation of Total, Average, and Marginal Revenue under Imperfect Competition

Given quantity and price data in the first two columns of Exhibit 3-7, total revenue, average revenue, and marginal revenue can be calculated for a firm that operates under imperfect competition.

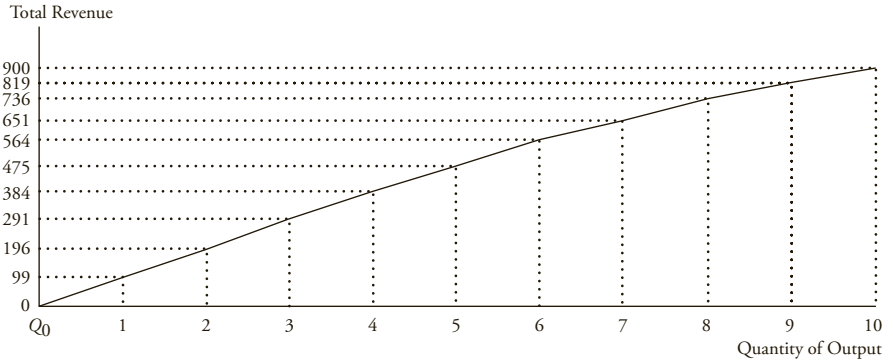
Describe how total revenue, average revenue, and marginal revenue change as quantity sold increases from 0 to 10 units.

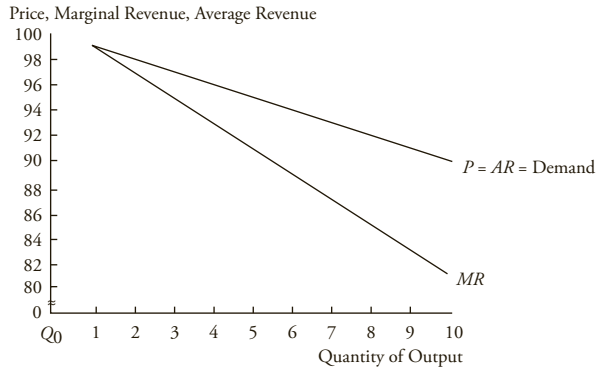
Solution: Total revenue increases with a greater quantity, but the rate of increase in *TR* (as measured by marginal revenue) declines as quantity increases. Average revenue and marginal revenue decrease when output increases, with *MR* falling faster than price and *AR*. Average revenue is equal to price at each quantity level. Exhibit 3-8 shows the relationships among the revenue variables presented in Exhibit 3-7.

EXHIBIT 3-7 Total, Average, and Marginal Revenue under Imperfect Competition

Quantity (*Q*)	Price (*P*)	Total Revenue (*TR*)	Average Revenue (*AR*)	Marginal Revenue (*MR*)
0	100	0	—	—
1	99	99	99	99
2	98	196	98	97
3	97	291	97	95
4	96	384	96	93
5	95	475	95	91
6	94	564	94	89
7	93	651	93	87
8	92	736	92	85
9	91	819	91	83
10	90	900	90	81

EXHIBIT 3-8 Total Revenue, Average Revenue, and Marginal Revenue for Exhibit 3-7 Data





The **quantity** or **quantity demanded** variable is the amount of the product that consumers are willing and able to buy at each price level. The quantity sold can be affected by the business through such activities as sales promotion, advertising, and competitive positioning of the product that would take place under the market model of imperfect competition. Under perfect competition, however, total quantity in the market is influenced strictly by price, whereas nonprice factors are not important. Once consumer preferences are established in the market, price determines the quantity demanded by buyers. Together, price and quantity constitute the firm's demand curve, which becomes the basis for calculating the total, average, and marginal revenue.

In Exhibit 3-4, **price** is the market price as established by the interactions of the market demand and supply factors. Since the firm is a price taker, price is fixed at 100 at all levels of output.

Total revenue (*TR*) is tabulated as price times the quantity of units sold. At one unit, *TR* is 100 (calculated as 100 × 1 unit); at 10 units it is 1,000 (calculated as 100 × 10 units). At zero quantity, obviously, total revenue is always zero. Under perfect competition, for each increment in quantity, total revenue increases by the price level, which is constant to the firm. This relationship is shown in Exhibit 3-4—where the increase in total revenue from one quantity to the next equals 100, which is equal to the price.

Average revenue (*AR*) is quantity sold divided into total revenue. The mathematical outcome of this calculation is simply the price that the firm receives in the market for selling a given quantity. For any firm that sells at a uniform price, average revenue will equal price. For example, *AR* at three units is 100 (calculated as 300 ÷ 3 units); at eight units it is also 100 (calculated as 800 ÷ 8 units).

Marginal revenue (*MR*) is the change in total revenue divided by the change in quantity sold; it is simply the additional revenue from selling one more unit. For example, in Exhibit 3-4, *MR* at four units is 100 [calculated as $(400 - 300) \div (4 - 3)$]; at nine units it is also 100 [calculated as $(900 - 800) \div (9 - 8)$]. In a competitive market in which price is constant to the individual firm regardless of the amount of output offered, marginal revenue is equal to average revenue, where both are the same as the market price. Reviewing the revenue data in Exhibit 3-4, price, average revenue, and marginal revenue are all equal to 100. In the case of imperfect competition, *MR* declines with greater output and is less than *AR* at any positive quantity level, as will become clear with Exhibit 3-7.

Exhibit 3-5 graphically displays the revenue data from Exhibit 3-4. For an individual firm operating in a market setting of perfect competition, *MR* equals *AR* and both are equal to a price that stays the same across all levels of output. Because price is fixed to the individual seller, the firm's demand curve is a horizontal line at the point where the market sets the price. In Exhibit 3-5, at a price of 100, $P_1 = MR_1 = AR_1 = \text{Demand}_1$. Marginal revenue, average revenue, and the firm's price remain constant until market demand and supply factors cause a change in price. For instance, if price increases to 200 because of an increase in market demand, the firm's demand curve shifts from Demand_1 to Demand_2 with corresponding increases in *MR* and *AR* as well. Total revenue increases from TR_1 to TR_2 when price increases from 100 to 200. At a price of 100, total revenue at 10 units is 1,000; however, at a price of 200, total revenue would be 2,000 for 10 units.

Exhibit 3-6 graphically illustrates the general shapes and relationships for *TR*, *AR*, and *MR* under imperfect competition. *MR* is positioned below the price and *AR* lines. *TR* peaks when *MR* equals zero at point Q_1.

3.1.2. Factors of Production

Revenue generation occurs when output is sold in the market. However, costs are incurred before revenue generation takes place as the firm purchases resources, or what are commonly

known as the factors of production, in order to produce a product or service that will be offered for sale to consumers. Factors of production, the inputs to the production of goods and services, include:

- *Land*, as in the site location of the business.
- *Labor*, which consists of the inputs of skilled and unskilled workers as well as the inputs of firms' managers.
- *Capital*, which in this context refers to *physical capital*—such tangible goods as equipment, tools, and buildings. Capital goods are distinguished as inputs to production that are themselves produced goods.
- *Materials*, which in this context refer to any goods the business buys as inputs to its production process.[1]

For example, a business that produces solid wood office desks needs to acquire lumber and hardware accessories as raw materials and hire workers to construct and assemble the desks using power tools and equipment. The factors of production are the inputs to the firm's process of producing and selling a product or service where the goal of the firm is to maximize profit by satisfying the demand of consumers. The types and quantities of resources or factors used in production, their respective prices, and how efficiently they are employed in the production process determine the cost component of the profit equation.

Clearly, in order to produce output, the firm needs to employ factors of production. While firms may use many different types of labor, capital, raw materials, and land, an analyst may find it more convenient to limit attention to a more simplified process in which only the two factors, capital and labor, are employed. The relationship between the flow of output and the two factors of production is called the **production function**, and it is represented generally as:

$$Q = f(K, L) \tag{3-5}$$

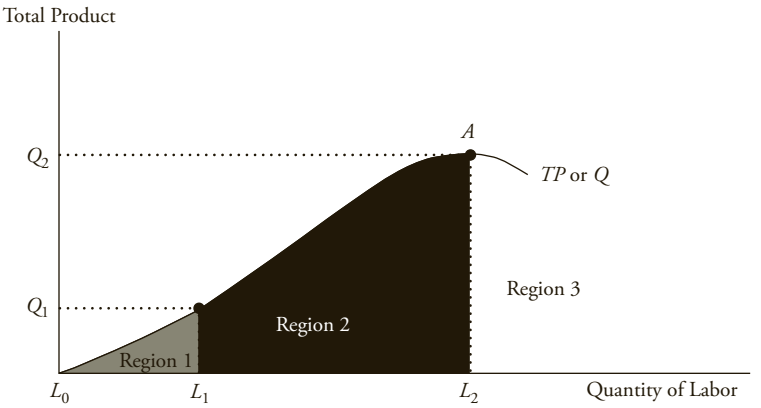
where Q is the quantity of output, K is capital, and L is labor. The inputs are subject to the constraint that $K \geq 0$ and $L \geq 0$. A more general production function is stated as:

$$Q = f(x_1, x_2, \ldots x_n) \tag{3-6}$$

where x_n represents the quantity of the nth input subject to $x_n \geq 0$ for n number of different inputs. Exhibit 3-9 illustrates the shape of a typical input–output relationship using labor (L) as the only variable input (all other input factors are held constant). The production function has three distinct regions where both the direction of change and the rate of change in total product (TP or Q, quantity of output) vary as production changes. Regions 1 and 2 have positive changes in TP as labor is added, but the change turns negative in Region 3. Moreover, in Region 1 (L_0 to L_1), TP is increasing at an increasing rate, typically because specialization allows laborers to become increasingly productive. In Region 2, however, (L_1 to L_2), TP is

[1]Because this factor may include such processed materials as steel and plastic that the firm purchases as inputs to production, the name *materials* was chosen in preference to another (traditional) name for this factor, *raw materials*. Candidates may encounter a number of variations in the classification and terminology for the factors of production.

EXHIBIT 3-9 A Firm's Production Function



increasing at a decreasing rate because capital is fixed, and labor experiences diminishing marginal returns. The firm would want to avoid Region 3 if at all possible because total product or quantity would be declining rather than increasing with additional input: There is so little capital per unit of labor that additional laborers would possibly get in each other's way. Point *A* is where *TP* is maximized.

EXAMPLE 3-3 Factors of Production

A group of business investors are in the process of forming a new enterprise that will manufacture shipping containers to be used in international trade.

1. What decisions about factors of production must the start-up firm make in beginning operations?
2. What objective should guide the firm in its purchase and use of the production factors?

Solution to 1: The entrepreneurs must decide where to locate the manufacturing facility in terms of an accessible site (land) and building (physical capital), what to use in the construction of the containers (materials), and what labor input to use.

Solution to 2: Overall, any decision involving the input factors should focus on how that decision affects costs, profitability, and risk such that shareholders' wealth is maximized.

3.1.3. Total, Average, Marginal, Fixed, and Variable Costs

Exhibit 3-10 shows the graphical relationships among total costs, total fixed cost, and total variable cost. The curve for total costs is a parallel shift of the total variable cost curve and

EXHIBIT 3-10 Total Costs, Total Variable Cost, and Total Fixed Cost

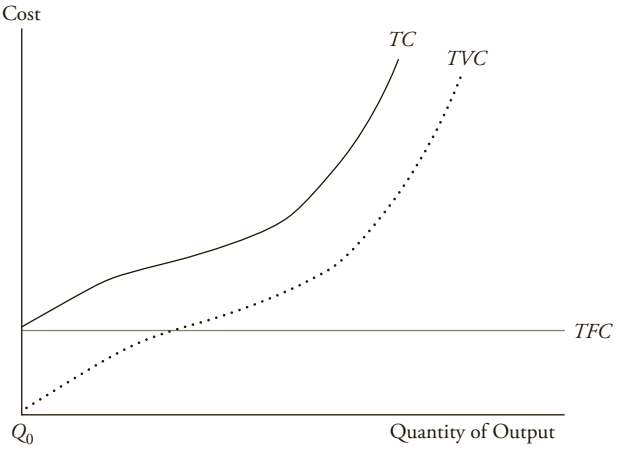
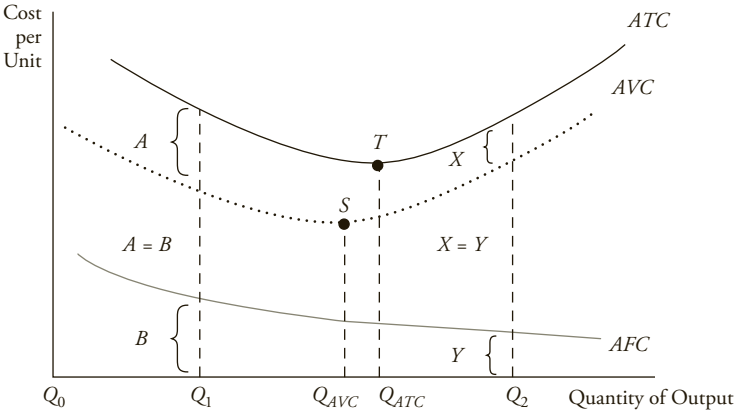


EXHIBIT 3-11 Average Total Cost, Average Variable Cost, and Average Fixed Cost



always lies above the total variable cost curve by the amount of total fixed cost. At zero production, total costs are equal to total fixed cost because total variable cost at this output level is zero.

Exhibit 3-11 shows the cost curve relationships among *ATC*, *AVC*, and *AFC* in the short run. (In the long run, the firm will have different *ATC*, *AVC*, and *AFC* cost curves when all inputs are variable, including technology, plant size, and physical capital.) The difference between *ATC* and *AVC* at any output quantity is the amount of *AFC*. For example, at Q_1 the distance between *ATC* and *AVC* is measured by the value of *A*, which equals the amount of fixed cost as measured by amount *B* at Q_1. Similarly, at Q_2, the distance between *ATC* and

AVC of *X* equals amount *Y* of *AFC*. The vertical distance between *ATC* and *AVC* is exactly equal to the height of *AFC* at each quantity. Both average total cost and average variable cost take on a bowl-shaped pattern in which each curve initially declines, reaches a minimum-cost output level, and then increases after that point. Point *S*, which corresponds to Q_{AVC}, is the minimum point on the *AVC* (such as two units in the next table). Similarly, point *T*, which corresponds to Q_{ATC}, is the minimum point on *ATC* (such as three units in the table). As shown in Exhibit 3-11, when output increases, average fixed cost declines as *AFC* approaches the horizontal quantity axis.

Exhibit 3-12 displays the cost curve relationships for *ATC*, *AVC*, and *MC* in the short run. The marginal cost curve intersects both the *ATC* and *AVC* at their respective minimum points. This occurs at points *S* and *T*, which correspond to Q_{AVC} and Q_{ATC}, respectively. Mathematically, when marginal cost is less than average variable cost, *AVC* will be decreasing. The opposite occurs when *MC* is greater than *AVC*. For example, in Exhibit 3-13, *AVC* begins to increase beyond two units, where *MC* exceeds *AVC*. The same relationship holds true for *MC* and *ATC*. Referring again to Exhibit 3-13, *ATC* declines up to three units when *MC* is less than *ATC*. After three units, *ATC* increases as *MC* exceeds *ATC*. Initially, the marginal cost curve declines, but at some point it begins to increase in reflection of an increasing rate of change in total costs as the firm produces more output. Point *R* (Exhibit 3-12), which corresponds to Q_{MC}, is the minimum point on the marginal cost curve.

Exhibit 3-13 shows an example of how total, average, and marginal costs are derived. Total costs are calculated by summing total fixed cost and total variable cost. Marginal cost is derived by taking the change in total costs as the quantity variable changes.

Exhibit 3-14 graphically displays the data for total costs, total variable cost, and total fixed cost from the table in Exhibit 3-13.

Total costs (*TC*) are the summation of all costs, where costs are classified according to fixed or variable. Total costs increase as the firm expands output and decrease when production is cut. The rate of increase in total costs declines up to a certain output level and, thereafter, accelerates as the firm gets closer to full utilization of capacity. The rate of change in total costs mirrors the rate of change in total variable cost. In Exhibit 3-13, *TC* at five units is

EXHIBIT 3-12 Average Total Cost, Average Variable Cost, and Marginal Cost

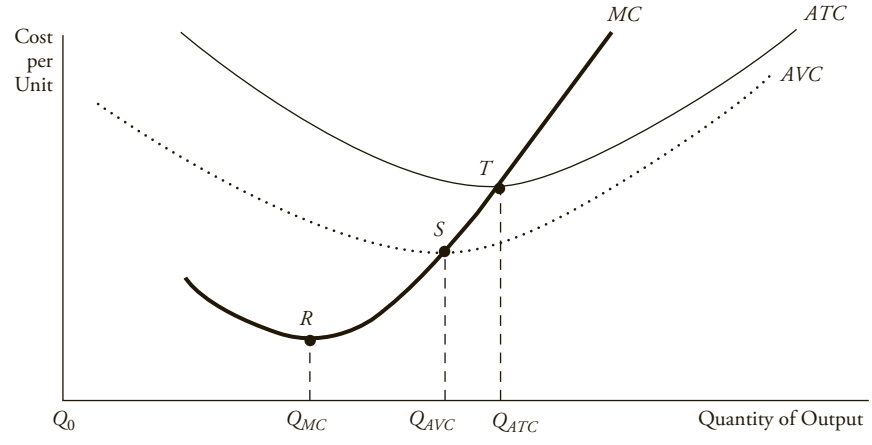
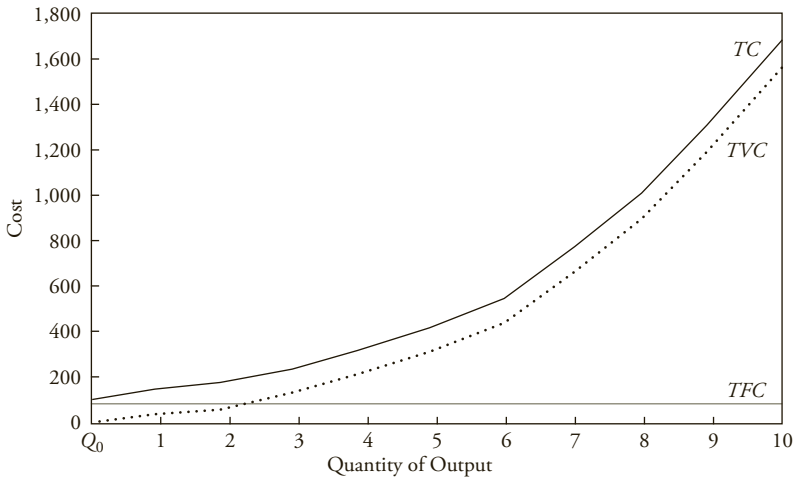


EXHIBIT 3-13 Total, Average, Marginal, Fixed, and Variable Costs

Quantity (Q)	Total Fixed Cost[a] (TFC)	Average Fixed Cost (AFC)	Total Variable Cost (TVC)	Average Variable Cost (AVC)	Total Costs (TC)	Average Total Cost (ATC)	Marginal Cost (MC)
0	100	—	0	—	100	—	—
1	100	100.0	50	50.0	150	150.0	50
2	100	50.0	75	37.5	175	87.5	25
3	100	33.3	125	41.7	225	75.0	50
4	100	25.0	210	52.5	310	77.5	85
5	100	20.0	300	60.0	400	80.0	90
6	100	16.7	450	75.0	550	91.7	150
7	100	14.3	650	92.9	750	107.1	200
8	100	12.5	900	112.5	1,000	125.0	250
9	100	11.1	1,200	133.3	1,300	144.4	300
10	100	10.0	1,550	155.0	1,650	165.0	350

[a]Includes all opportunity costs.

EXHIBIT 3-14 Total Costs, Total Variable Cost, and Total Fixed Cost for Exhibit 3-13 Data



400—of which 300 is variable cost and 100 is fixed cost. At 10 units, total costs are 1,650, which is the sum of 1,550 in variable cost and 100 in fixed cost.

Total fixed cost (TFC) is the summation of all expenses that do not change when production varies. It can be a sunk or unavoidable cost that a firm has to cover regardless of whether it produces anything at all, or it can be a cost that stays the same over a range of production but can change to another constant level when production moves outside of that

range. The latter is referred to as a **quasi-fixed cost**, although it remains categorized as part of *TFC*. Examples of fixed costs are debt service, real estate lease agreements, and rental contracts. Quasi-fixed cost examples would be certain utilities and administrative salaries that could be avoided or be lower when output is zero but would assume higher constant values over different production ranges. Normal profit is considered to be a fixed cost because it is a return required by investors on their equity capital regardless of output level. At zero output, total costs are always equal to the amount of total fixed cost that is incurred at this production point. In Exhibit 3-13, total fixed cost remains at 100 throughout the entire production range.

Other fixed costs evolve primarily from investments in such fixed assets as real estate, production facilities, and equipment. As a firm grows in size, fixed asset expansion occurs along with a related increase in fixed cost. However, fixed cost cannot be arbitrarily cut when production declines. Regardless of the volume of output, an investment in a given level of fixed assets locks the firm into a certain amount of fixed cost that is used to finance the physical capital base, technology, and other capital assets. When a firm downsizes, the last expense to be cut is usually fixed cost.

Total variable cost (*TVC*), which is the summation of all variable expenses, has a direct relationship with quantity. When quantity increases, total variable cost increases; total variable cost declines when quantity decreases. At zero production, total variable cost is always zero. Variable cost examples are payments for labor, raw materials, and supplies. As indicated earlier, total costs mirror total variable cost, with the difference being a constant fixed cost. The change in total variable cost (which defines marginal cost) declines up to a certain output point and then increases as production approaches capacity limits. In Exhibit 3-13, total variable cost increases with an increase in quantity. However, the change from one to two units is 25, calculated as $(75 - 50)$; the change from 9 to 10 units is 350, calculated as $(1{,}550 - 1{,}200)$.

Another approach to calculating total variable cost is to determine the variable cost per unit of output and multiply this cost figure by the number of production units. Per-unit variable cost is the cost of producing each unit exclusive of any fixed cost allocation to production units. One can assign variable cost individually to units or derive an average variable cost per unit.

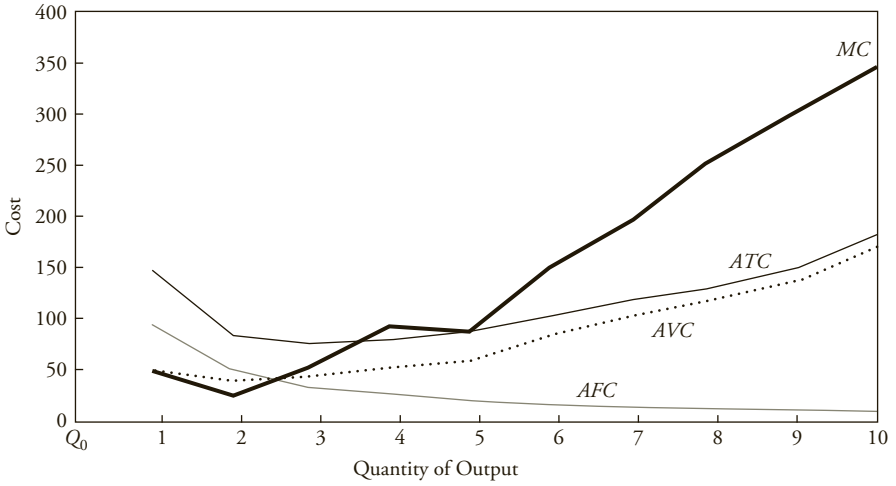
Whenever a firm initiates a downsizing, retrenchment, or defensive strategy, variable cost is the first to be considered for reduction, given its variability with output. However, variable cost is reducible only so far because all firms have to maintain a minimum amount of labor and other variable resources to function effectively.

Exhibit 3-15 illustrates the relationships among marginal cost, average total cost, average variable cost, and average fixed cost for the data presented in Exhibit 3-13.

Dividing total fixed cost by quantity yields **average fixed cost** (*AFC*), which decreases throughout the production span. A declining average fixed cost reflects spreading a constant cost over more and more production units. At high production volumes, *AFC* may be so low that it is a small proportion of average total cost. In Exhibit 3-13, *AFC* declines from 100 at one unit to 20 at five units, and then to 10 at an output level of 10 units.

Average variable cost (*AVC*) is derived by dividing total variable cost by quantity. For example, average variable cost at five units is $(300 \div 5)$ or 60. Over an initial range of production, average variable cost declines and then reaches a minimum point. Thereafter, cost increases as the firm utilizes more of its production capacity. This higher cost results primarily from production constraints imposed by the fixed assets at higher volume levels. The minimum point on the *AVC* coincides with the lowest average variable cost. However, the minimum point on the *AVC* does not correspond to the least-cost quantity for average total cost.

EXHIBIT 3-15 Average Total Cost, Average Variable Cost, Average Fixed Cost, and Marginal Cost for Exhibit 3-13 Data



In Exhibit 3-13, average variable cost is minimized at two units, whereas average total cost is the lowest at three units.

Average total cost (*ATC*) is calculated by dividing total costs by quantity or by summing average fixed cost and average variable cost. For instance, in Exhibit 3-13, at eight units *ATC* is 125, calculated as (1,000 ÷ 8) or (*AFC* + *AVC* = 12.5 + 112.5). Average total cost is often referenced as per-unit cost and is frequently called average cost. The minimum point on the average total cost curve defines the output level that has the least cost. The cost-minimizing behavior of the firm would dictate operating at the minimum point on its *ATC* curve. However, the quantity that maximizes profit (such as Q_3 in the next diagram) may not correspond to the ATC-minimum point. The minimum point on the *ATC* curve is consistent with maximizing profit per unit, but it is not necessarily consistent with maximizing total profit. In Exhibit 3-13, the least-cost point of production is three units; *ATC* is 75, derived as (225 ÷ 3) or (33.3 + 41.7). Any other production level results in a higher *ATC*.

Marginal cost (*MC*) is the change in total cost divided by the change in quantity. Marginal cost also can be calculated by taking the change in total variable cost and dividing by the change in quantity. It represents the cost of producing an additional unit. For example, at nine units marginal cost is 300, calculated as (1,300 − 1,000) ÷ (9 − 8). Marginal cost follows a J-shaped pattern whereby cost initially declines but turns higher at some point in reflection of rising costs at higher production volumes. In Exhibit 3-13, *MC* is the lowest at two units of output with a value of 25, derived as (175 − 150) ÷ (2 − 1).

Exhibit 3-17 displays the firm's supply curve, shutdown point, and breakeven level of operation under perfect competition in the short run. The firm's **short-run supply curve** is the bold section of the marginal cost curve that lies above the minimum point (point *A*) on the average variable cost curve. If the firm operates below this point (for example between *C* and *A*), it shuts down because of its inability to cover variable costs in full. Between points *A* and *B*, the firm can operate in the short run because it is meeting variable cost payments even

EXAMPLE 3-4 Calculation and Interpretation of Total, Average, Marginal, Fixed, and Variable Costs

The first three columns of Exhibit 3-16 display data on quantity, total fixed cost, and total variable cost, which are used to calculate total costs, average fixed cost, average variable cost, average total cost, and marginal cost. Interpret the results for total, average, marginal, fixed, and variable costs.

EXHIBIT 3-16 Total, Average, Marginal, Fixed, and Variable Costs

Q	*TFC*[a]	*TVC*	*AFC*	*AVC*	*TC*	*ATC*	*MC*
0	5,000	0	—	—	5,000	—	—
1	5,000	2,000	5,000.0	2,000	7,000	7,000.0	2,000
2	5,000	3,800	2,500.0	1,900	8,800	4,400.0	1,800
3	5,000	5,400	1,666.7	1,800	10,400	3,466.7	1,600
4	5,000	8,000	1,250.0	2,000	13,000	3,250.0	2,600
5	5,000	11,000	1,000.0	2,200	16,000	3,200.0	3,000
6	5,000	15,000	833.3	2,500	20,000	3,333.3	4,000
7	5,000	21,000	714.3	3,000	26,000	3,714.3	6,000
8	5,000	28,800	625.0	3,600	33,800	4,225.0	7,800
9	5,000	38,700	555.6	4,300	43,700	4,855.6	9,900
10	5,000	51,000	500.0	5,100	56,000	5,600.0	12,300

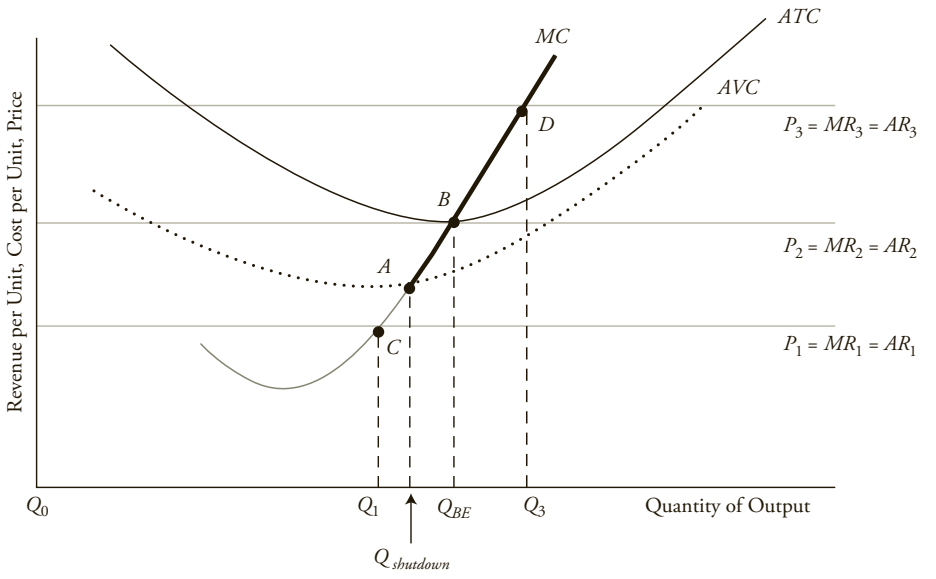
[a]Includes all opportunity costs.

Solution: Total fixed cost remains unchanged at 5,000 throughout the entire production range, while average fixed cost continuously declines from 5,000 at one unit to 500 by 10 units. Both average variable cost and marginal cost initially decline and then reach their lowest level at three units, with costs of 1,800 and 1,600, respectively. Beyond three units, both average variable cost and marginal cost increase, indicating that the cost of production rises with greater output. The least-cost point for average total cost is 3,200 at five units. At zero output, total costs are 5,000, equal to the amount of total fixed cost.

though it is unable to cover all of its fixed costs. In the long run, however, the firm is not able to survive if fixed costs are not completely covered. Any operating point above point *B* (the minimum point on *ATC*), such as point *D*, generates an economic profit.

A firm's **shutdown point** occurs when average revenue is less than average variable cost (any output below $Q_{shutdown}$), which corresponds to point *A* in Exhibit 3-17. Shutdown is defined as a situation in which the firm stops production but still confronts the payment of fixed costs in the short run as a business entity. In the short run, a business is capable

EXHIBIT 3-17 A Firm's Short-Run Supply Curve, Breakeven Point, and Shutdown Point under Perfect Competition



of operating in a loss situation as long as it covers its variable costs even though it is not earning sufficient revenue to cover all fixed cost obligations. If variable costs cannot be covered in the short run ($P < AVC$), the firm will shut down operations and simply absorb the unavoidable fixed costs. This problem occurs at output Q_1, which corresponds to point C where price is less than average variable cost. However, in the long run, to remain in business, the price must cover all costs. Therefore, in the long run, at any price below the breakeven point, the firm will exit the market; it will no longer participate in the market. Point D, which corresponds to output Q_3, is a position where economic profit occurs because price is greater than ATC.

In the case of perfect competition, the **breakeven point** is the quantity where price, average revenue, and marginal revenue equal average total cost. It is also defined as the quantity where total revenue equals total costs. Firms strive to reach initial breakeven as soon as possible to avoid start-up losses for any extended period of time. When businesses are first established, there is an initial period when losses occur at low quantity levels. In Exhibit 3-17, the breakeven quantity occurs at output Q_{BE}, which corresponds to point B where price is tangent to the minimum point on the ATC. (Keep in mind that normal profit as an implicit cost is included in ATC as a fixed cost.)

Exhibit 3-18 shows the breakeven point under perfect competition using the total revenue–total cost approach. Actually, there are two breakeven points—lower (point E) and upper (point F). Below point E, the firm is losing money (economic losses), and above that point is the region of profitability (shaded area) that extends to the upper breakeven point. Within this profit area, a specific quantity (Q_{max}) maximizes profit as the largest difference between TR and TC. Point F is where the firm leaves the profit region and incurs economic losses again. This second region of economic losses develops when the firm's production begins to reach the limits of physical capacity, resulting in diminished productivity and an acceleration of costs. Obviously, the firm would not produce beyond Q_{max}, because that is the optimal production point that maximizes profit.

EXHIBIT 3-18 A Firm's Breakeven Points Using Total Revenue and Total Costs under Perfect Competition

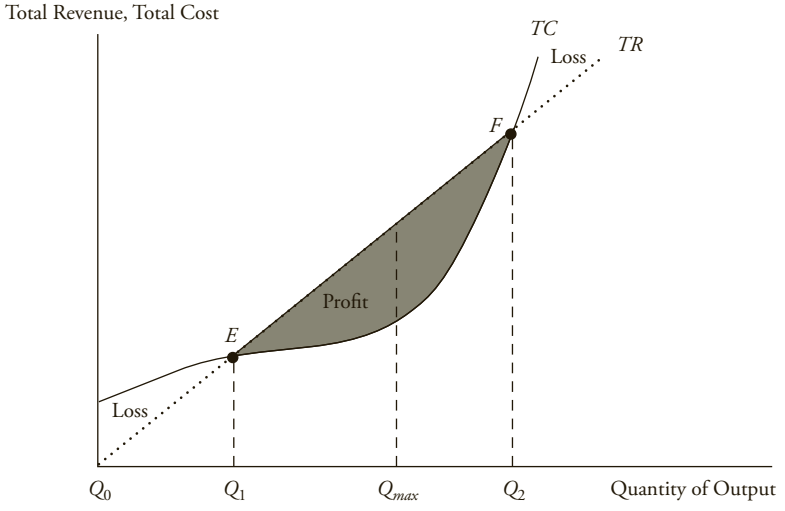
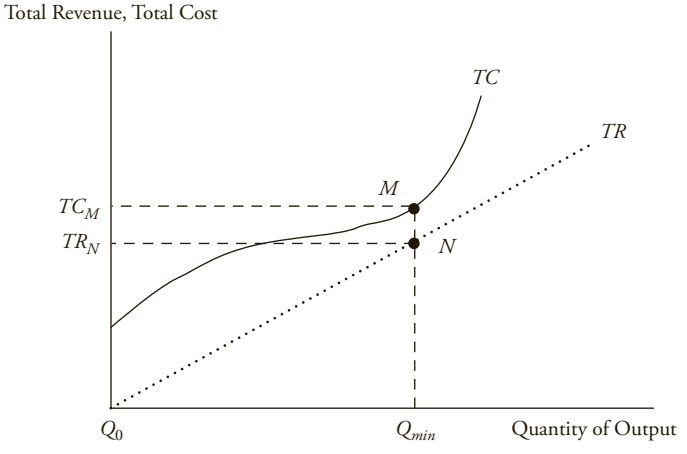


EXHIBIT 3-19 Loss Minimization Using Total Revenue and Total Costs



Breakeven points, profit regions, and economic loss ranges are influenced by demand and supply conditions, which change frequently according to the market behavior of consumers and firms. A high initial breakeven point is riskier than a low point because it takes a larger volume and, usually, a longer time to reach. However, at higher output levels it yields more return in compensation for this greater risk.

In the case where *TC* exceeds *TR*, as shown in Exhibit 3-19, the firm will want to minimize the economic loss (as long as $TR > TVC$), which is defined as the smallest difference between *TC* and *TR*. This occurs at Q_{min}, where the economic loss is calculated as ($TC_M - TR_N$) on the vertical axis.

EXAMPLE 3-5 Breakeven Analysis and Profit Maximization When the Firm Faces a Negatively Sloped Demand under Imperfect Competition

The following revenue and cost information for a future period is presented in Exhibit 3-20 for WR International, a newly formed corporation that engages in the manufacturing of low-cost, prefabricated dwelling units for urban housing markets in emerging economies. (Note that quantity increments are in blocks of 10 for a 250 change in price.) The firm has few competitors in a market setting of imperfect competition.

1. How many units must WR International sell to initially break even?
2. Where is the region of profitability?
3. At what point will the firm maximize profit? At what points are there economic losses?

EXHIBIT 3-20 Breakeven and Profit Analysis

Quantity (*Q*)	Price (*P*)	Total Revenue (*TR*)	Total Costs (*TC*)[a]	Profit
0	10,000	0	100,000	(100,000)
10	9,750	97,500	170,000	(72,500)
20	9,500	190,000	240,000	(50,000)
30	9,250	277,500	300,000	(22,500)
40	9,000	360,000	360,000	0
50	8,750	437,500	420,000	17,500
60	8,500	510,000	480,000	30,000
70	8,250	577,500	550,000	27,500
80	8,000	640,000	640,000	0
90	7,750	697,500	710,000	(12,500)
100	7,500	750,000	800,000	(50,000)

[a]Includes all opportunity costs.

Solution to 1: WR International will initially break even at 40 units of production, where *TR* and *TC* equal 360,000.

Solution to 2: The region of profitability will range from 40 to 80 units. Any production quantity of less than 40 units and any quantity greater than 80 will result in an economic loss.

Solution to 3: Maximum profit of 30,000 will occur at 60 units. Lower profit will occur at any output level that is higher or lower than 60 units. From zero quantity to 40 units and for quantities beyond 80 units, economic losses occur.

EXHIBIT 3-21 Short-Run and Long-Run Decisions to Operate, Shut Down Product, or Exit the Market

Revenue–Cost Relationship	Short-Run Decision	Long-Term Decision
$TR \geq TC$	Stay in market	Stay in market
$TR > TVC$ but $TR < TFC + TVC$	Stay in market	Exit market
$TR < TVC$	Shut down production to zero	Exit market

Given the relationships among total revenue, total variable cost, and total fixed cost, Exhibit 3-21 summarizes the decisions to operate, shut down production, or exit the market in both the short run and the long run. As previously discussed, the firm must cover variable cost before fixed cost. In the short run, if total revenue cannot cover total variable cost, the firm shuts down production to minimize loss, which would equal the amount of fixed cost. If total variable cost exceeds total revenue in the long run, the firm will exit the market as a business entity to avoid the loss associated with fixed cost at zero production. By terminating business operations through market exit, investors escape the erosion in their equity capital from economic losses. When total revenue is enough to cover total variable cost but not all of total fixed cost, the firm can survive in the short run but will be unable to maintain financial solvency in the long run.

EXAMPLE 3-6 Shutdown Analysis

For the most recent financial reporting period, a business domiciled in Ecuador (which recognizes the U.S. dollar as an official currency) has revenue of $2 million and total costs of $2.5 million, which are or can be broken down into total fixed cost of $1 million and total variable cost of $1.5 million. The net loss on the firm's income statement is reported as $500,000 (ignoring tax implications). In prior periods, the firm had reported profits on its operations.

1. What decision should the firm make regarding operations over the short term?
2. What decision should the firm make regarding operations over the long term?
3. Assume the same business scenario except that revenue is now $1.3 million, which creates a net loss of $1.2 million. What decision should the firm make regarding operations in this case?

Solution to 1: In the short run, the firm is able to cover all of its total variable cost but only half of its $1 million in total fixed cost. If the business ceases to operate, its loss is $1 million, the amount of total fixed cost, whereas the net loss by operating is minimized at $500,000. The firm should attempt to operate by negotiating special arrangements with creditors to buy time to return operations back to profitability.

Solution to 2: If the revenue shortfall is expected to persist over time, the firm should cease operations, liquidate assets, and pay debts to the extent possible. Any residual for shareholders would decrease the longer the firm is allowed to operate unprofitably.

Solution to 3: The firm would minimize loss at $1 million of total fixed cost by shutting down compared with continuing to do business where the loss is $1.2 million. Shareholders will save $200,000 in equity value by pursuing this option. Unquestionably, the business would have a rather short life expectancy if this loss situation were to continue.

When evaluating profitability, particularly of start-up firms and businesses using turnaround strategies, analysts should consider highlighting breakeven and shutdown points in their financial research. Identifying the unit sales levels where the firm enters or leaves the production range for profitability and where the firm can no longer function as a viable business entity provides invaluable insight to investment decisions.

3.1.4. Output Optimization and Maximization of Profit

Profit maximization occurs when:

- The difference between total revenue (*TR*) and total costs (*TC*) is the greatest.
- Marginal revenue (*MR*) equals marginal cost (*MC*).
- The revenue value of the output from the last unit of input employed equals the cost of employing that input unit (as later developed in Equation 3-12).

All three approaches derive the same profit-maximizing output level. In the first approach, a firm starts by forecasting unit sales, which becomes the basis for estimates of future revenue and production costs. By comparing predicted total revenue to predicted total costs for different output levels, the firm targets the quantity that yields the greatest profit. When using the marginal revenue–marginal cost approach, the firm compares the change in predicted total revenue (*MR*) with the change in predicted total costs (*MC*) by unit of output. If *MR* exceeds *MC*, total profit is increased by producing more units because each successive unit adds more to total revenue than it does to total costs. If *MC* is greater than *MR*, total profit is decreased when additional units are produced. The point of profit maximization occurs where *MR* equals *MC*. The third method compares the estimated cost of each unit of input to that input's contribution with projected total revenue. If the increase in projected total revenue coming from the input unit exceeds its cost, a contribution to total profit is evident. In turn, this justifies further employment of that input. On the other hand, if the increase in projected total revenue does not cover the input unit's cost, total profit is diminished. Profit maximization based on the employment of inputs occurs where the next input unit for each type of resource used no longer makes any contribution to total profit.

Combining revenue and cost data from Exhibits 3-4 and 3-13, Exhibit 3-22 demonstrates the derivation of the optimal output level that maximizes profit for a firm under perfect competition. Profit is calculated as the difference between total revenue and total costs. At zero

EXHIBIT 3-22 Profit Maximization under Perfect Competition

Quantity (*Q*)	Price (*P*)	Total Revenue (*TR*)	Total Costs (*TC*)[a]	Profit (*P*)	Marginal Revenue (*MR*)	Marginal Cost (*MC*)
0	100	0	100	(100)	—	—
1	100	100	150	(50)	100	50
2	100	200	175	25	100	25
3	100	300	225	75	100	50
4	100	400	310	90	100	85
5	100	500	400	100	100	90
6	100	600	550	50	100	150
7	100	700	750	(50)	100	200
8	100	800	1,000	(200)	100	250
9	100	900	1,300	(400)	100	300
10	100	1,000	1,650	(650)	100	350

[a]Includes all opportunity costs.

production, an economic loss of 100 occurs, which is equivalent to total fixed cost. Upon initial production, the firm incurs an economic loss of 50 on the first unit but breaks even by unit 2. The region of profitability ranges from two to six units. Within this domain, total profit is maximized in the amount of 100 at five units of output. No other quantity level yields a higher profit. At this five-unit level, marginal revenue exceeds marginal cost. But at unit 6, marginal revenue is less than marginal cost, which results in a lower profit because unit 6 costs more to produce than it generates in revenue. Unit 6 costs 150 to produce but contributes only 100 to total revenue, which yields a 50 loss on that unit. As a result, profit drops from 100 to 50. At unit 7 and beyond, the firm begins to lose money again as it passes the upper breakeven mark and enters a second economic loss zone.

Exhibits 3-23 and 3-24 display the data from Exhibit 3-22 to illustrate profit maximization under perfect competition using the $(TR - TC)$ and $(MR = MC)$ approaches. Exhibit 3-24 highlights profit maximization based on comparing how much each unit of output costs to produce (*MC*) to how much each unit contributes to revenue (*MR*). Each unit up to and including unit 5 contributes to profit in that each unit's marginal revenue exceeds its marginal cost. Starting at unit 6 and thereafter, the marginal revenue for each unit is less than the marginal cost. This results in a reduction in profit. Profit maximization occurs where *MR* equals *MC*. In this case, the optimal decision for the firm using a comparison of *MR* and *MC* is to produce five units.[2]

[2]Marginal analysis is a common and valuable optimization tool that is used to determine the point of profit maximization and the optimal employment of resources. It is often said that the firm makes decisions "on the margin."

EXHIBIT 3-23 Profit Maximization Using Total Revenue and Total Costs from Exhibit 3-22

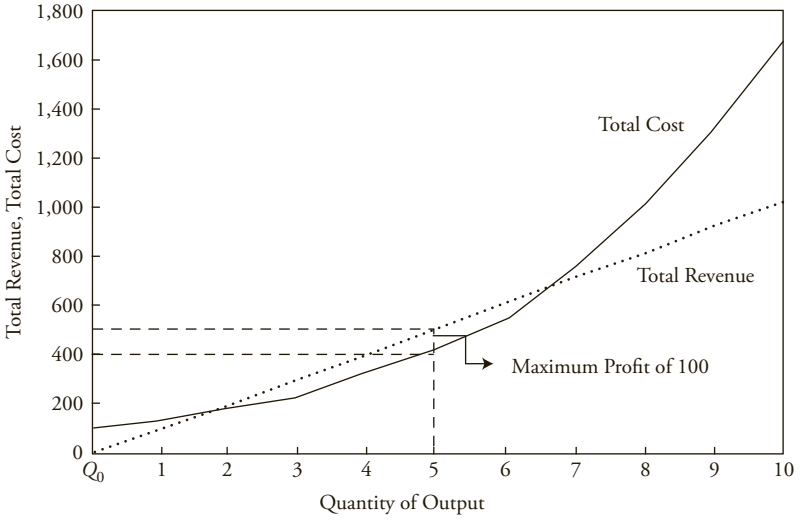
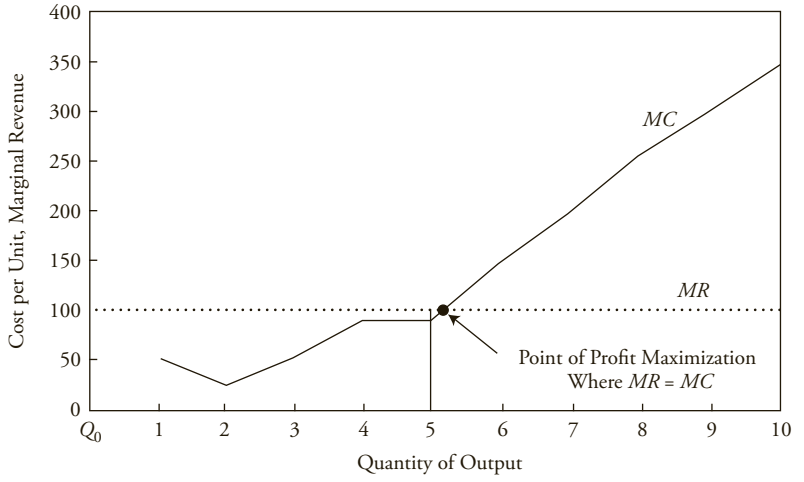


EXHIBIT 3-24 Profit Maximization Using Marginal Revenue and Marginal Cost from Exhibit 3-22



It should be noted that under imperfect competition, the firm faces a negatively sloped demand curve. As the firm offers a greater quantity to the market, price decreases. In contrast, a firm under perfect competition has an insignificant share of the market and is able to sell more without impacting market price. Obviously, the type of market structure in which a firm operates as a seller has an impact on the firm's profit in terms of the price received when output levels vary.

EXAMPLE 3-7 Profit Maximization and the Breakeven Point under Imperfect Competition

Exhibit 3-25 shows revenue and cost data for a firm that operates under the market structure of imperfect competition.

1. At what point does the firm break even over its production range in the short run?
2. What is the quantity that maximizes profit given total revenue and total costs?
3. Comparing marginal revenue and marginal cost, determine the quantity that maximizes profit.

EXHIBIT 3-25 Breakeven and Profit Analysis under Imperfect Competition

Q	*P*	*TR*	*TC*[a]	Profit	*MR*	*MC*
0	1,000	0	550	(550)	—	—
1	995	995	1,000	(5)	995	450
2	990	1,980	1,500	480	985	500
3	985	2,955	2,100	855	975	600
4	980	3,920	2,800	1,120	965	700
5	975	4,875	3,600	1,275	955	800
6	970	5,820	4,600	1,220	945	1,000
7	965	6,755	5,800	955	935	1,200
8	960	7,680	7,200	480	925	1,400
9	955	8,595	8,800	(205)	915	1,600
10	950	9,500	10,800	(1,300)	905	2,000

[a]Includes all opportunity costs.

Solution to 1: The breakeven point occurs between unit 1 and unit 2, where profit increases from (5) to 480.

Solution to 2: At an output level of five units, the firm maximizes profit in the amount of 1,275, calculated as the difference between *TR* of 4,875 and *TC* of 3,600.

Solution to 3: Profit maximization occurs at five units, where *MR* of 955 exceeds *MC* of 800, which yields a profit contribution of 155. However, at six units, *MR* of 945 is less than the *MC* of 1,000, resulting in a loss of 55 and a reduction in profit from 1,275 to 1,220.

Exhibit 3-26 summarizes the ($TR - TC$) and ($MR = MC$) profit-maximization approaches for firms operating under perfect competition. (Profit maximization using inputs is discussed in Section 3.2.2.)

Profit acts as an efficient allocator of equity capital to investment opportunities whereby shareholders' wealth is increased. Equity flows from low-return business investments to high-return business investments as it seeks the greatest return potential on a risk-adjusted basis. Basic economic theory describes how consumer choice voiced through the price mechanism in competitive markets directs resources to their most efficient use according to what consumers need and want. In the end, it is profitability—which evolves from the interactions of demand and supply factors in product and resource markets—that decides where financial capital is employed.

3.1.5. Economies of Scale and Diseconomies of Scale

Rational behavior dictates that the firm select an operating size or scale that maximizes profit over any time frame. The time frame for the firm can be separated into the short run and the long run based on the ability of the firm to adjust the quantities of the fixed resources it employs. The short run is defined as a time period in which at least one of the factors of production is fixed. The most likely inputs to be held constant in defining the short run are technology, physical capital, and plant size. Usually, a firm cannot change these inputs in a relatively short period of time, given the inflexible nature of their use. The long run is defined as a time period in which all factors of production are variable, including technology, physical capital, and plant size. Additionally, in the long run, firms can enter or exit the market based on decisions regarding profitability. The long run is often referred to as the **planning horizon** in which the firm can choose the short-run position or optimal operating size that maximizes profit over time.

The time required for long-run adjustments varies by industry. For example, the long run for a small business using very little in the way of technology and physical capital may be less than a year. In contrast, for a capital-intensive firm, the long run may be more than a decade. However, given enough time, all production factors are variable, which allows the firm to choose an operating size or plant capacity based on different technologies and physical capital. In this regard, costs and profits will differ between the short run and the long run.

EXHIBIT 3-26 Summary of Profit Maximization and Loss Minimization under Perfect Competition

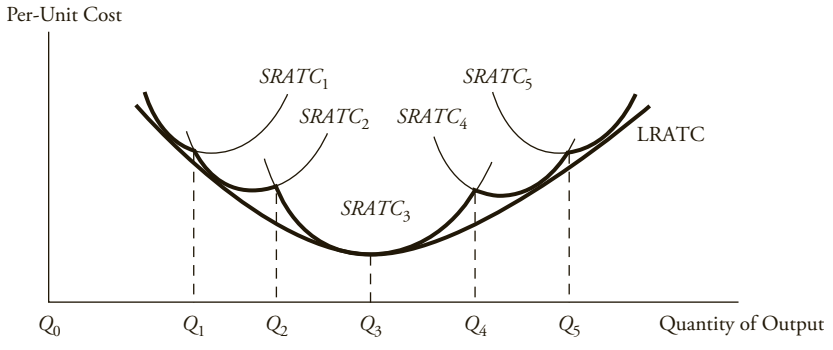
Revenue–Cost Relationship	Actions by Firm
$TR = TC$ and $MR > MC$	Firm is operating at lower breakeven point; increase Q to enter profit territory.
$TR \geq TC$ and $MR = MC$	Firm is at maximum profit level; make no change in Q.
$TR < TC$ and $TR \geq TVC$ but ($TR - TVC$) $< TFC$ (covering TVC but not TFC)	Find level of Q that minimizes loss in the short run; work toward finding a profitable Q in the long run; exit market if losses continue in the long run.
$TR < TVC$ (not covering TVC in full)	Shut down in the short run; exit market in the long run.
$TR = TC$ and $MR < MC$	Firm is operating at upper breakeven point; decrease Q to enter profit territory.

The fixed-input constraint in the short run along with input prices establish the firm's **short-run average total cost curve** (SRATC). This defines what the per-unit cost will be for any quantity in the short run. The SRATC and the demand for the firm's product determine short-run profit. The selection of technology, physical capital, and plant size is a key determinant of the short-run cost curve for the firm. As the firm switches to newer technologies and physical capital, a corresponding change in short-run costs occurs. In the long run, a firm has the opportunity for greater profit potential based on the ability to lower its costs through choices of more efficient technology and physical capital and a wider selection of production capacities.

Exhibit 3-27 displays the **long-run average total cost curve** (LRATC), which is derived from the short-run average total cost curves that are available to the firm.[3] The business has a choice of five technology–physical capital options and plant capacities over the long run, each with its own short-run cost curve. The LRATC consists of sections of these individual short-run cost curves. For example, from zero production to Q_1 output, $SRATC_1$ yields the lowest per-unit cost. Between Q_1 and Q_2, the lowest cost per unit is attainable with $SRATC_2$, which represents a larger production capacity. $SRATC_3$ and $SRATC_4$ would provide for the lowest average cost over time for output levels of Q_2 to Q_4 and Q_4 to Q_5, respectively. For any output greater than Q_5, $SRATC_5$ becomes the preferred curve for minimizing average total cost. Tangentially connecting all of the least-cost SRATC segments by way of an envelope curve creates the LRATC. (Assuming an unlimited number of possible technologies, plant sizes, and physical capital combinations—and therefore a theoretically unlimited number of SRATCs—the LRATC becomes a smooth curve rather than a segmented one as indicated by the bold segments of the five SRATCs in Exhibit 3-27.) The LRATC shows the lowest cost per unit at which output can be produced over a long period of time when the firm is able to make technology, plant size, and physical capital adjustments. If the same technologies and physical capital are available and adaptable to every firm in the industry, then all firms would have the same LRATC. However, firms could be at different positions on this homogeneous LRATC, depending on their operating size that is based on output.

Over the long run, as a business expands output, it can utilize more efficient technology and physical capital and take advantage of other factors to lower the costs of production. This

EXHIBIT 3-27 Long-Run Average Total Cost Curve



[3]Some writers use short-run average cost (SRAC) and long-run average cost (LRAC) in the same sense as SRATC and LRATC, respectively.

development is referred to as **economies of scale** or **increasing returns to scale** as a firm moves to lower cost structures when it grows in size. Output increases by a larger proportion than the increase in inputs. The opposite effect can result after a certain volume level at which the business faces higher costs as it expands in size. This outcome is called **diseconomies of scale** or **decreasing returns to scale**, where the firm becomes less efficient with size. In this case, output increases by a smaller proportion than the increase in inputs. Diseconomies of scale often result from the firm becoming too large to be managed efficiently even though better technology can be increasing productivity within the business. Both economies and diseconomies of scale can occur at the same time; the impact on long-run average total cost depends on which dominates. If economies of scale dominate, LRATC decreases with increases in output; the reverse holds true when diseconomies of scale prevail.

Referring back to Exhibit 3-27, economies of scale occur from Q_0 (zero production) to output level Q_3, where Q_3 is the cost-minimizing level of output for $SRATC_3$. It is evident throughout this production range that per-unit costs decline as the firm produces more. Over the production range of Q_3 to Q_5, diseconomies of scale are occurring as per-unit costs increase when the firm expands output. Under perfect competition, given the five SRATC selections that are available to the firm throughout the production range Q_0 to Q_5, $SRATC_3$ is the optimal technology, plant capacity, and physical capital choice, with Q_3 being the target production size for the firm that would minimize cost over the long term.

Perfect competition forces the firm to operate at the minimum point on the LRATC because market price will be established at this level over the long run. If the firm is not operating at this least-cost point, its long-term viability will be threatened. The minimum point on the LRATC is referred to as the **minimum efficient scale** (MES). The MES is the optimal firm size under perfect competition over the long run at which the firm can achieve cost competitiveness.

As the firm grows in size, economies of scale and a lower average total cost can result from the following factors:

- Division of labor and management in a large firm with numerous workers, where each worker can specialize in one task rather than performing many duties, as in the case of a small business (accordingly, workers in a large firm become more proficient at their jobs).
- Being able to afford more expensive, yet more efficient equipment and to adapt the latest in technology that increases productivity.
- Effectively reducing waste and lowering costs through marketable by-products, less energy consumption, and enhanced quality control.
- Better use of market information and knowledge for more effective managerial decision making.
- Discounted prices on resources when buying in larger quantities.

A classic example of a business that realizes economies of scale through greater physical capital investment is the electric utility. By expanding output capacity to accommodate a larger customer base, the utility company's per-unit cost will decline. Economies of scale help to explain why electric utilities have naturally evolved from localized entities to regional and multiregional enterprises. Wal-Mart, the world's largest retailer, is an example of a business that uses bulk purchasing power to obtain deep discounts from suppliers to keep costs and prices low. Wal-Mart also utilizes the latest in technology to monitor point-of-sale transactions to have timely market information to respond to changes in customer buying behavior. This leads to economies of scale through lower distribution and inventory costs.

The factors that can lead to diseconomies of scale, inefficiencies, and rising costs when a firm increases in size include:

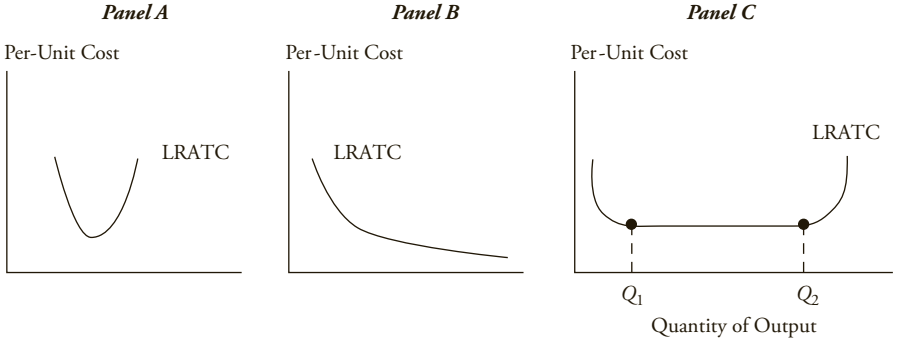
- Becoming so large that it cannot be properly managed.
- Overlap and duplication of business functions and product lines.
- Higher resource prices because of supply constraints when buying inputs in large quantities.

General Motors (GM) is an example of a business that has realized diseconomies of scale by becoming too large. Scale diseconomies have occurred through product overlap and duplication (i.e., similar or even identical automobile models), where the fixed cost for these models is not spread over a large volume of output. (Recently, the company has decided to discontinue various low-volume product models that overlapped with other models.) GM has numerous manufacturing plants throughout the world and sells vehicles in over a hundred countries. Given this geographical dispersion in production and sales, the company has had communication and management coordination problems, which have resulted in higher costs. Also, GM has had significantly higher labor costs than its competitors. By being the largest producer in the market, it has been a target of labor unions for higher compensation and benefits packages relative to other firms.

Strategically, when a firm is operating in the economies of scale region, expanding production capacity will increase the firm's competitiveness through lower costs. Firm expansion is often facilitated with a growth or business combination (i.e., merger or acquisition) strategy. However, when a business is producing in the area of diseconomies of scale, the objective is to downsize to reduce costs and become more competitive. From an investment perspective, a firm operating at the minimum point of the industry LRATC under perfect competition should be valued higher than a firm that is not producing at this least-cost quantity.

The LRATC can take various forms given the development of new technology and growth prospects for an industry over the long term. Exhibit 3-28 displays examples of different average total cost curves that firms can realize over the long run. Panel A shows that scale economies dissipate rapidly at low output levels. This implies that a firm with a low volume of output can be more cost competitive than a firm that is producing a high output volume. Panel B indicates a lower and lower average cost over time as firm size increases. The larger the business, the more competitive it is and the greater its potential investment value. Finally, Panel C shows the case of **constant returns to scale** (i.e., output increases by the same proportion as the increase in inputs) over the range of production from Q_1 to Q_2, indicating

EXHIBIT 3-28 Types of Long-Run Average Total Cost Curves

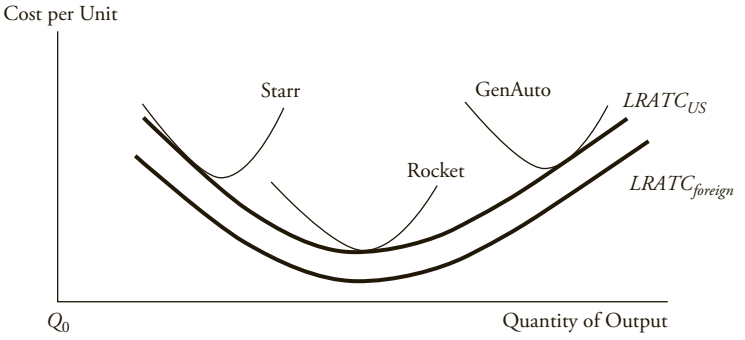


EXAMPLE 3-8 Long-Run Average Total Cost Curve

Exhibit 3-29 displays the long-run average total cost curve ($LRATC_{US}$) and the short-run average total cost curves for three hypothetical U.S.-based automobile manufacturers—Starr Vehicles (Starr), Rocket Sports Cars (Rocket), and General Auto (GenAuto). The long-run average total cost curve ($LRATC_{foreign}$) for foreign-owned automobile companies that compete in the U.S. auto market is also indicated in the graph. (The market structure implicit in the exhibit is imperfect competition.)

To what extent are the cost relationships depicted in Exhibit 3-29 useful for an economic and financial analysis of the three U.S.-based auto firms?

EXHIBIT 3-29 Long-Run Average Total Cost Curve and the Short-Run Average Total Cost Curves for Starr Vehicles (Starr), Rocket Sports Cars (Rocket), and General Auto (GenAuto)



Solution: First, it is observable that the foreign auto companies have a lower LRATC than the U.S. automobile manufacturers. This competitive position places the U.S. firms at a cost and possible pricing disadvantage in the market with the potential to lose market share to the lower-cost foreign competitors. Second, only Rocket operates at the minimum point of the $LRATC_{US}$, whereas GenAuto is situated in the region of diseconomies of scale and Starr is positioned in the economies of scale portion of the curve. To become more efficient and competitive, GenAuto needs to downsize and restructure, which means moving down the $LRATC_{US}$ curve to a smaller, yet lower-cost production volume. In contrast, Starr has to grow in size to become more efficient and competitive by lowering per-unit costs.

From a long-term investment prospective and given its cost advantage, Rocket has the potential to create more investment value relative to GenAuto and Starr. Over the long run, if GenAuto and Starr can lower their average total costs, they will become more attractive to investors. However, if any or all of the three U.S. auto companies cannot match the cost competitiveness of the foreign firms, they may be driven from the market. In the long run, the lower-cost foreign automakers pose a severe competitive challenge to the survival of the U.S. manufacturers and their ability to maintain and grow shareholders' wealth.

that size does not give a firm a competitive edge over another firm within this range. In other words, a firm that is producing the smaller output Q_1 has the same long-run average total cost as a firm producing the higher output Q_2.

3.1.6. Profit Maximization in the Short Run and the Long Run

No matter the time span, the firm's supply behavior centers on the objective of profit maximization. In the short term, when technology and physical capital are fixed, maximum profit (or minimal loss) is determined where marginal cost equals marginal revenue (points *A* and *B* in Exhibit 3-30). Cases of profit maximization and loss minimization are illustrated in Exhibit 3-30 for a firm operating under perfect competition in the short run. In Panel A, the firm realizes economic profit because *TR* is greater than *TC* and price exceeds SRATC in the production range of Q_1 to Q_2. Q_{max} is the output level that maximizes economic profit. Panel B shows the case of loss minimization because *TC* exceeds *TR* and SRATC is above the price level. Q_{min} yields the least loss of all possible production quantities. Note that in this case, the short-run loss is still less than fixed cost, so the firm should continue operating in the short run.

Exhibit 3-31 illustrates long-run profit maximization under perfect competition given the long-run average total cost curve when economies of scale occur. In the long run under perfect competition, the firm will operate at the minimum efficient scale point on its long-run average total cost curve. This least-cost point is illustrated in Exhibit 3-31 as point *E* at output level Q_E. In comparison to the point of minimum efficient scale, any other output quantity results in a higher cost.

In the short run, given $SRATC_1$ and P_1, the firm is making only normal profit because price equals average total cost at point *A*. By realizing economies of scale in the long run, the firm can move down the LRATC to $SRATC_2$ and produce Q_E. If the firm still receives P_1, economic profit is forthcoming at Q_E in the amount of $(B - E)$ per unit. However, economic

EXHIBIT 3-30 Profit Maximization and Loss Minimization in the Short Run under Perfect Competition

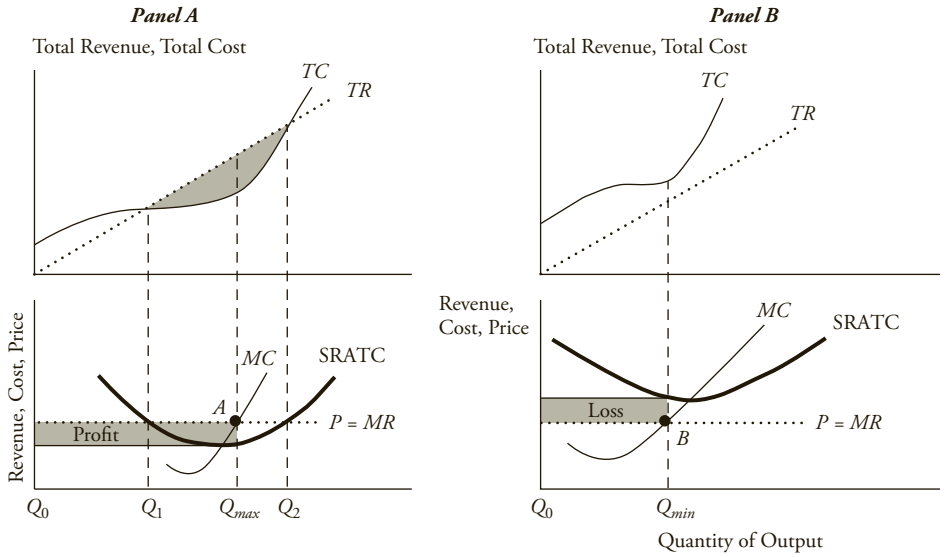


EXHIBIT 3-31 Long-Run Profit Maximization and Minimum Efficient Scale under Perfect Competition

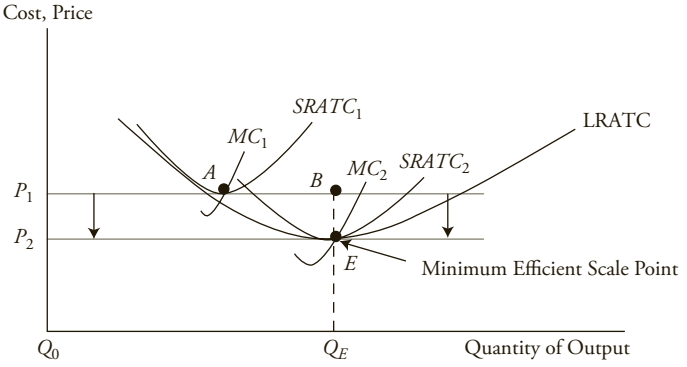
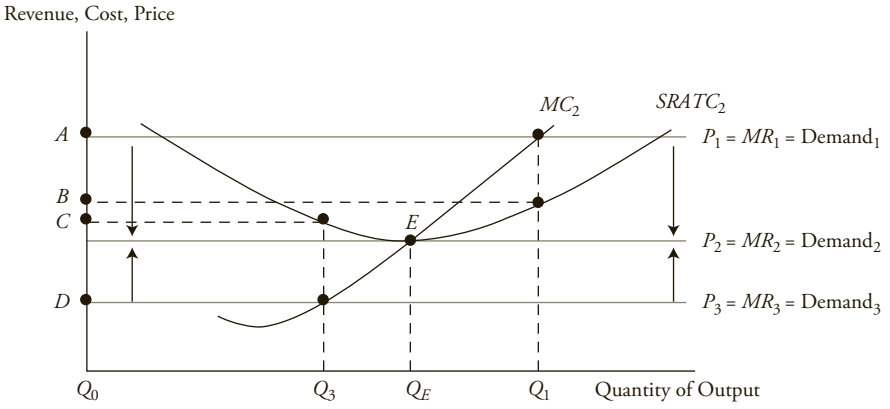


EXHIBIT 3-32 Profit Maximization and Loss Minimization in the Long Run under Perfect Competition



profit with no barriers to entry under perfect competition leads to more competitors, a greater market supply, and, subsequently, a lower price in the long run. The price to the firm will decline to P_2, and economic profit will disappear with the long-run equilibrium for the firm occurring at point E, the minimum efficient scale. At point E, the firm is making only normal profit because in the long run under perfect competition, economic profit is zero.

Exhibit 3-32 illustrates profit maximization and loss minimization in the long run when market prices change for a firm that is operating in a market of perfect competition at the minimum efficient scale point on its LRATC. ($SRATC_2$, MC_2, Q_E, and point E are the same in both Exhibit 3-31 and Exhibit 3-32.) Although point E represents the lowest production cost, the quantity that maximizes profit or minimizes loss is determined where marginal revenue equals marginal cost. (Under perfect competition, price equals marginal revenue.) If price is at P_1 (which equals MR_1), the firm will produce Q_1 and accrue economic profit of $(A - B)$ per unit in the short run because price is greater than average total cost. In the long

run, economic profit attracts new competitors that drive price down, resulting in zero economic profit. Profitability declines to the level at P_2, where price is tangent to average total cost at point E. If price is at P_3 (which equals MR_3), the firm will produce Q_3 and realize an economic loss of $(C - D)$ per unit in the short run because average total cost is greater than price. In the long run, firms will exit the market; as a result, price rises to P_2, eliminating economic losses. Again, profitability for the firm returns to the normal level at point E, where price matches average total cost. The long-term equilibrium for the firm occurs at point E, which corresponds to Demand_2, MR_2, a price of P_2, and an output level of Q_E.

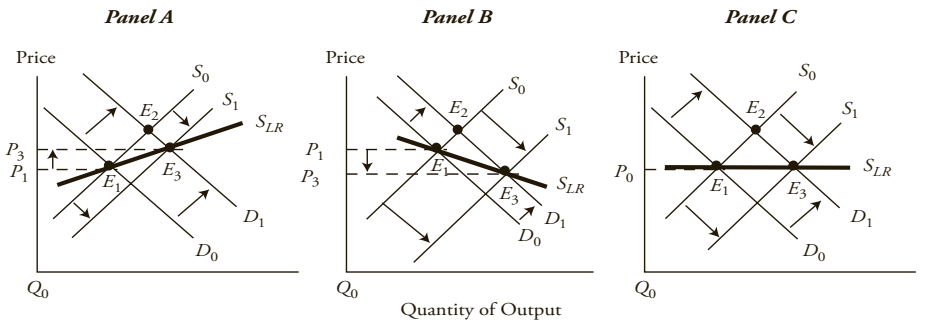
3.1.7. The Long-Run Industry Supply Curve and What It Means for the Firm

The **long-run industry supply curve** shows the relationship between quantities supplied and output prices for an industry when firms are able to enter or exit the industry in response to the level of short-term economic profit (i.e., perfect competition) and when changes in industry output influence resource prices over the long run. Exhibit 3-33 illustrates three types of long-run supply curves based on increasing costs, decreasing costs, and constant costs to firms competing under perfect competition.

An **increasing-cost industry** exists when prices and costs are higher when industry output is increased in the long run. This is demonstrated in Panel A. Assuming zero economic profit at E_1, when demand increases from D_0 to D_1, price rises and economic profit results at E_2 in the short run. Over the long term in response to this economic profit, new competitors will enter the industry and existing firms will expand output, resulting in an increase in supply from S_0 to S_1 and a long-term equilibrium at E_3. If the increase in demand for resources from this output expansion leads to higher prices for some or all inputs, the industry as a whole will face higher production costs and charge a higher price for output. As indicated by S_{LR} in Panel A, the long-run supply curve for the industry will have a positive slope over the long run. The firm in an increasing-cost industry will experience higher resource costs, so market price must rise in order to cover these costs. The petroleum, coal, and natural gas industries are prime examples of increasing-cost industries, where the supply response to long-run demand growth results in higher output prices because of the rising costs of energy production.

Panel B shows the case of a **decreasing-cost industry**, where the supply increase from S_0 to S_1 leads to a lower price for output in the market. Firms are able to charge a lower price because of a reduction in their resource costs. Decreasing costs can evolve from technological advances, producer efficiencies that come from a larger firm size, and economies of scale of resource suppliers (i.e., lower resource prices) that are passed on to resource buyers when industry output expands. The long-run supply curve for the industry will have a negative slope,

EXHIBIT 3-33 Long-Run Supply Curves for the Firm



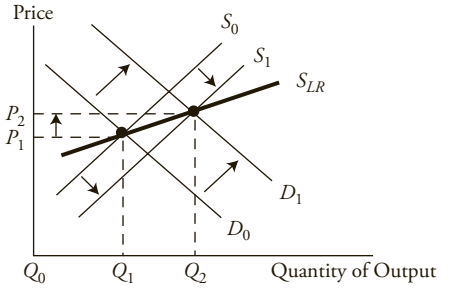
as displayed by S_{LR} in Panel B. As a result, the firm in a decreasing-cost industry will experience lower resource costs and can then charge a lower price.[4] Possible examples of decreasing-cost industries are semiconductors and personal computers, where lower production costs and the rapid growth in demand over the past decade have led to substantially lower prices.

EXAMPLE 3-9 A Firm Operating in an Increasing-Cost Industry

Mirco Industries is a global manufacturer of outdoor recreational equipment in a market setting of easy entry and price competition. Company forecasts of total market demand for outdoor recreational products over the long run indicate robust growth in sales as families allocate more time to outdoor and leisure activities. This scenario looks promising for Mirco's earnings and shareholders' value. To assist Mirco in its assessment of this future scenario, industry analysts have presented Exhibit 3-34 to illustrate industry costs and the market supply curve over the long run.

Using Exhibit 3-34, what information would be of value to Mirco in identifying future production cost and price under a market growth scenario?

EXHIBIT 3-34 Industry Costs and the Long-Run Industry Supply Curve



Note: Market demand increases from D_0 to D_1. The market responds with an increase in supply from S_0 to S_1. In the long run, the price of P_2 will be higher than the original price of P_1.

Solution: As indicated by the upward slope of the long-run industry supply curve, Mirco will experience an increase in production costs over the long run because of higher resource prices when the industry expands production and new firms enter the industry in response to an increase in market demand. To cover the higher production costs, Mirco will ultimately charge the higher market price of P_2 relative to the current price of P_1.

[4]The individual firm's supply curve is still upward sloping even though the industry's long-run supply curve is negatively sloped. This means that in the long run, the firm's supply curve shifts to the right when industry costs decrease. This results in a lower price charged by the firm for each quantity.

In some cases, firms in the industry will experience no change in resource costs and output prices over the long run. This type of industry is known as a **constant-cost industry**. This is displayed in Panel C, where the long-run supply curve (S_{LR}) for the industry is a horizontal line indicating a constant price level when the industry increases output.

3.2. Productivity

In general terms, **productivity** is defined as the average output per unit of input. Any production factor can be used as the input variable. However, it has been a common practice to use the labor resource as the basis for measuring productivity. In this regard, productivity is based on the number of workers used or the number of hours of work performed. In many cases, labor is easier to quantify than are the other types of resources used in production. Therfore, productivity is typically stated as output per worker or output per labor hour.

Why is productivity important? Cost minimization and profit maximization behavior dictate that the firm strives to maximize productivity—that is, produce the most output per unit of input or produce any given level of output with the least amount of inputs. A firm that lags behind the industry in productivity is at a competitive disadvantage and, as a result, is most likely to face decreases in future earnings and shareholders' wealth. An increase in productivity lowers production costs, which leads to greater profitability and investment value. Furthermore, productivity benefits (e.g., increased profitability) can be fully or partially distributed to other stakeholders of the business, such as consumers in the form of lower prices and employees in the form of enhanced compensation. Transferring some or all of the productivity rewards to others besides equity holders creates synergies that benefit shareholders over time.

The benefits from increased productivity are:

- Lower business costs, which translate into increased profitability.
- An increase in the market value of equity and shareholders' wealth resulting from an increase in profit.
- An increase in worker rewards, which motivates further productivity increases from labor.

Undoubtedly, increases in productivity reinforce and strengthen the competitive position of the firm over the long run. A fundamental analysis of a company should examine the firm's commitment to productivity enhancements and the degree to which productivity is integrated into the competitive nature of the industry or market. In some cases, productivity is not only an important promoter of growth in firm value over the long term, but it is also the key factor for economic survival. A business that lags the market in terms of productivity often finds itself less competitive and at the same time confronting profit erosion and deterioration in shareholders' wealth. Whenever productivity is a consideration in the equity valuation of the firm, the first step for the analyst is to define measures of productivity. Typical productivity measures for the firm are based on the concepts of total product, average product, and marginal product of labor.

3.2.1. Total, Average, and Marginal Product of Labor

When measuring a firm's operating efficiency, it is easier and more practical to use a single resource factor as the input variable rather than a bundle of the different resources that the firm uses in producing units of output. As discussed in the previous section, labor is typically

the input that is the most identifiable and calculable for measuring productivity. However, any input that is not difficult to quantify can be used. An example will illustrate the practicality of using a single-factor input, such as labor, to evaluate the firm's output performance. A business that manually assembles widgets has 50 workers, one production facility, and an assortment of equipment and hand tools. The firm would like to assess its productivity when it utilizes these three types of input factors to produce widgets. In this case, the most appropriate method is to use labor as the input factor for determining productivity, because the firm uses a variety of physical capital and only one plant building.

To illustrate the concepts of total product, average product, and marginal product, labor is used as the input variable. Exhibit 3-35 provides a summary of definitions and tabulations for these three concepts.

Measured on the basis of the labor input, **total product** (*TP* or *Q*) is defined as the aggregate sum of production for the firm during a time period. As a measure of productivity, total product provides superficial information as to how effective and efficient the firm is in terms of producing output. For instance, three firms—Company A, Company B, and Company C—that comprise the entire industry have total output levels of 100,000 units, 180,000 units, and 200,000 units, respectively. Obviously, Company C dominates the market with a 41.7 percent share, followed by Company B's 37.5 percent share and Company A's 20.8 percent portion of the market. This information says little about how efficient each firm is in generating its total output level. Total product only provides an insight into the firm's production volume relative to the industry; it does not show how efficient the firm is in producing its output.

Average product (*AP*) measures the productivity of inputs on average and is calculated by dividing total product by the total number of units for a given input that is used to generate that output. Average product is usually measured on the basis of the labor input. It is a representative or overall measure of labor's productivity: some workers are more productive than average, and others are less productive than average.

Given the aforementioned production levels for the three firms, Company A employs 100 workers, and Company B and Company C utilize 200 and 250 workers, respectively. Calculating average product of labor for each of the three firms yields the following productivity results: Company A → 1,000 units of output per worker, Company B → 900 units per worker, and Company C → 800 units per worker. It is apparent that Company A is the most efficient firm, although it has the lowest share of the total market. Company C has the largest portion

EXHIBIT 3-35 Definitions and Calculations for Total, Marginal, and Average Product of Labor

Term	Calculation
Total product	Sum of the output from all inputs during a time period; usually illustrated as the total output (*TP* or *Q*) using labor (*L*).
Average product	Total product divided by the quantity of a given input; *AP* is measured as total product divided by the number of workers used at that output level; ($TP \div L$) or ($Q \div L$).
Marginal product	The amount of additional output resulting from using one more unit of input assuming other inputs are fixed; *MP* is measured by taking the difference in total product and dividing by the change in the quantity of labor; ($\Delta TP \div \Delta L$) or ($\Delta Q \div \Delta L$).

of the total market, but it is the least efficient of the three. Given that Company A can maintain its productivity advantage over the long run, it will be positioned to generate the greatest return on investment through lower costs and higher profit outcomes relative to the other firms in the market.

Marginal product (*MP*), also known as marginal return, measures the productivity of each unit of input and is calculated by taking the difference in total product from adding another unit of input (assuming other resource quantities are held constant). Typically, it is measured in terms of labor's performance; thereby, it is a gauge of productivity of the individual additional worker rather than an average across all workers.

Exhibit 3-36 provides a numerical illustration for total, average, and marginal products of labor.

Total product increases as the firm adds labor until worker 7; at that point total production declines by 70 units. Obviously, the firm does not want to employ any worker who has negative productivity. In this case, no more than six workers are considered for employment with the firm.

At an employment level of five workers, *AP* and *MP* are 80 units (400 ÷ 5) and 40 units [(400 − 360) ÷ (5 − 4)], respectively. The productivity of the fifth worker is 40 units, while the average productivity for all five workers is 80 units, twice that of worker 5.

A firm has a choice of using total product, average product, marginal product, or some combination of the three to measure productivity. Total product does not provide an in-depth view of a firm's state of efficiency. It is simply an indication of a firm's output volume and potential market share. Therefore, average product and marginal product are better gauges of a firm's productivity because both can reveal competitive advantage through production efficiency. However, individual worker productivity is not easily measurable when workers perform tasks collectively. In this case, average product is the preferred measure of productivity performance.

3.2.2. Marginal Returns and Productivity

Referring to the marginal product column in Exhibit 3-36, worker 2 has a higher output of 110 units compared with worker 1, who produces 100 units; there is an increase in return when employees are added to the production process. This economic phenomenon is known

EXHIBIT 3-36 Total, Average, and Marginal Product of Labor

Labor (L)	Total Product (TP_L)	Average Product (AP_L)	Marginal Product (MP_L)
0	0	—	—
1	100	100	100
2	210	105	110
3	300	100	90
4	360	90	60
5	400	80	40
6	420	70	20
7	350	50	(70)

as **increasing marginal returns**, where the marginal product of a resource increases as additional units of that input are employed. However, successive workers beyond number 2 have lower and lower marginal product to the point where the last worker has a negative return. This observation is called the **law of diminishing returns**. Diminishing returns can lead to a negative marginal product, as evidenced with worker 7. There is no question that a firm does not want to employ a worker or input that has a negative impact on total output.

Initially, a firm can experience increasing returns from adding labor to the production process because of the concepts of specialization and division of labor. At first, by having too few workers relative to total physical capital, the understaffing situation requires employees to multitask and share duties. As more workers are added, employees can specialize, become more adept at their individual functions, and realize an increase in marginal productivity. But after a certain output level, the law of diminishing returns becomes evident.

Assuming all workers are of equal quality and motivation, the decline in marginal product is related to the short run, where at least one resource (typically plant size, physical capital, or technology) is fixed. When more and more workers are added to a fixed plant size/technology/physical capital base, the marginal return of the labor factor eventually decreases because the fixed input restricts the output potential of additional workers. One way of understanding the law of diminishing returns is to void the principle and assume that the concept of increasing returns lasts indefinitely. As more workers are added, or when any input is increased, the marginal output continuously increases. At some point, the world's food supply could be grown on one hectare of land or all new automobiles could be manufactured in one factory. Physically, the law of increasing returns is not possible in perpetuity, even though it can clearly be evident in the early stages of production.

Another element resulting in diminishing returns is the quality of labor itself. In the previous discussion, it was assumed all workers were of equal ability. However, that assumption may not be entirely valid when the firm's supply of labor has varying degrees of human capital. In that case, the business would want to employ the most productive workers first; then, as the firm's labor demand increased, less productive workers would be hired. When the firm does not have access to an adequate supply of homogeneous human capital, or for that matter any resource, diminishing marginal product occurs at some point.

The data provided in Exhibit 3-38 show productivity changes for various U.S. industries over the period from 2000 to 2007. The coal mining and newspaper sectors have several years of negative changes in productivity, which do not reinforce prospects for long-term growth in profitability. Declines in productivity raise production costs and reduce profit. For the most part, the other industries have solid productivity increases from year to year, even though in

EXAMPLE 3-10 Calculation and Interpretation of Total, Average, and Marginal Product

Average product and marginal product can be calculated on the basis of the production relationship between the number of machines and total product, as indicated in the first two columns of Exhibit 3-37.

1. Interpret the results for total, average, and marginal product.
2. Indicate where increasing marginal returns change to diminishing marginal returns.

EXHIBIT 3-37 Total, Average, and Marginal Product

Machines (K)	Total Product (TP_K)	Average Product (AP_K)	Marginal Product (MP_K)
0	0	—	—
1	1,000	1,000	1,000
2	2,500	1,250	1,500
3	4,500	1,500	2,000
4	6,400	1,600	1,900
5	7,400	1,480	1,000
6	7,500	1,250	100
7	7,000	1,000	(500)

Solution to 1: Total product increases to six machines, where it tops out at 7,500. Because total product declines from machine 6 to machine 7, the marginal product for machine 7 is negative 500 units. Average product peaks at 1,600 units with four machines.

Solution to 2: Increasing returns are evident up to machine 3, where marginal product equals 2,000 units of output. Beyond machine 3, decreasing returns develop because MP_K declines when more machines are added to the production process.

EXHIBIT 3-38 Productivity Changes for Selected Sectors, 2000–2007

Sector	NAICS[a]	2000	2001	2002	2003	2004	2005	2006	2007
Coal mining	212,100	4.9%	−1.3%	−2.3%	1.5%	0.0%	−4.9%	−7.5%	1.2%
Newspaper publishers	511,110	5.5	−4.3	−0.6	5.1	−5.6	2.4	4.0	−1.8
Auto	336,100	−10.6	0.3	14.5	12.0	1.1	4.6	10.2	4.8
Commercial banking	522,110	3.9	−2.3	4.3	4.5	5.5	1.3	2.9	0.9
Merchandise stores	452,000	5.9	3.8	3.5	6.0	2.8	3.2	3.3	0.5
Air transportation	481,000	1.9	−5.3	9.9	10.2	12.7	7.6	5.1	1.8

[a]North American Industry Classification System.
Note: Productivity is defined by the U.S. Bureau of Labor Statistics as output per worker-hour.
Source: U.S. Bureau of Labor Statistics, "Productivity and Costs by Industry: Annual Rates of Change."

some cases the change is volatile. On a trend basis, productivity increases appear to have peaked in the period 2002–2004 and then edged downward during the latter part of the period. Declining productivity makes a firm or industry less competitive over time; however, any adverse impact on profitability stemming from lower or negative productivity may be offset by rising demand for the product.

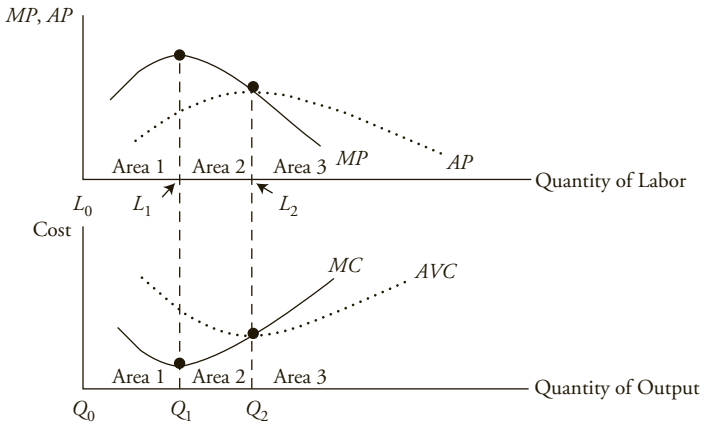
Productivity is a key element in the determination of costs and profit to the firm, especially over the long term. Although productivity can fluctuate widely in the short run (as indicated in Exhibit 3-38) for a variety of reasons, secular (long-term) patterns in output per unit of labor denote more meaningful relationships among productivity, costs, profits, and the competitive status of the firm with respect to the industry. To summarize, the analyst should study the productivity levels of the firm over the long run and do an evaluation of how the firm's efficiency compares with the industry standard. A firm that lags the industry in productivity may find itself at a competitive disadvantage with the end result of profit erosion and negative implications for shareholders' wealth. Once evident, productivity issues cause the firm's market value of equity to be discounted.

As previously discussed, a major determinant of the cost component of the profit equation is the degree of efficiency with which the firm uses resources in producing output as defined by the firm's production function. Given the relationship between output and inputs, marginal product (*MP*) and average product (*AP*) form the basis for marginal cost (*MC*) and average variable cost (*AVC*). Actually, *MC* and *AVC* are respective mirror images of *MP* and *AP*. Exhibit 3-39 illustrates this relationship in the short run by showing three areas of interest.

Area 1 shows an increasing *MP* from L_0 to L_1. The increases in *MP* result in declining marginal costs from Q_0 to Q_1. As *MP* or productivity peaks at L_1, *MC* is minimized at Q_1. Diminishing marginal returns take over in areas 2 and 3, where a decreasing marginal product results in higher marginal costs. Not only does *MP* impact *MC*, but the shape of the *AVC* also is based on the pattern of *AP*. At L_2, *AP* is maximized, while its corresponding output level of Q_2 is consistent with the minimum position on the *AVC* curve. Note that when *MP* is greater than *AP*, *AP* is increasing; when *MP* is less than *AP*, *AP* is declining. A similar relationship holds true for *MC* and *AVC*. When *MC* is less than *AVC*, *AVC* is decreasing; the opposite occurs when *MC* is greater than *AVC*. In area 3, *AP* is declining, which creates an upturn in the *AVC* curve.

Technology, quality of human and physical capital, and managerial ability are key factors in determining the production function relationship between output and inputs. The firm's production function establishes what productivity is in terms of *TP*, *MP*, and *AP*. In turn, productivity significantly influences total, marginal, and average costs to the firm, and costs

EXHIBIT 3-39 Relationship of Average Product and Marginal Product to Average Variable Cost and Marginal Cost in the Short Run



directly impact profit. Obviously, what happens at the production level in terms of productivity impacts the cost level and profitability.

Because revenue, costs, and profit are measured in monetary terms, the productivity of the different input factors requires comparison on a similar basis. In this regard, the firm wants to maximize output per monetary unit of input cost. This goal is denoted by the following expression:

$$MP_{input}/P_{input} \tag{3-7}$$

where MP_{input} is the marginal product of the input factor and P_{input} is the price of that factor (i.e., resource cost).

When using a combination of resources, a least-cost optimization formula is constructed as follows:

$$\frac{MP_1}{\text{Price of input 1}} = \cdots = \frac{MP_n}{\text{Price of input } n} \tag{3-8}$$

where the firm utilizes n different resources. Using a two-factor production function consisting of labor and physical capital, Equation 3-9 best illustrates this rule of least cost:

$$\frac{MP_L}{P_L} = \frac{MP_K}{P_K} \tag{3-9}$$

where MP_L and MP_K are the marginal products of labor and physical capital, respectively. P_L is the price of labor or the wage rate, and P_K is the price of physical capital. For example, if MP_L/P_L equals 2 and MP_K/P_K is 4, physical capital yields twice the output per monetary unit of input cost versus labor. It is obvious that the firm will want to use physical capital over labor in producing additional output because it provides more productivity on an equivalent cost basis. However, as more physical capital is employed, the firm's MP of capital declines because the law of diminishing returns impacts production. Physical capital is added until its ratio of MP per monetary unit of input cost matches that of labor: $MP_K/P_K = MP_L/P_L = 2$.[5] At this point, both inputs are added when expanding output until their ratios differ. When their ratios diverge, the input with the higher ratio will be employed over the other lower ratio input when the firm increases production.

Equations 3-7, 3-8, and 3-9 derive the physical output per monetary unit of input cost. However, to determine the profit-maximizing utilization level of an input, the firm must measure the revenue value of the input's MP and then compare this figure with the cost of the input. The following equations represent this relationship:

$$\text{Marginal product} \times \text{Product price} = \text{Price of the input} \tag{3-10}$$

$$\text{Marginal revenue product} = \text{Price of the input} \tag{3-11}$$

Marginal revenue product (*MRP*) is calculated as the MP of an input unit times the price of the product. This term measures the value of the input to the firm in terms of what

[5]This assumes that MP_L is independent of physical capital, K. However, as more K is used, MP_L could actually increase because labor will become more productive when using more physical capital. In this case, $MP_K/P_K = MP_L/P_L$ at some point between 2 and 4.

EXAMPLE 3-11 Determining the Optimal Input Combination

Canadian Global Electronic Corp. (CGEC) uses three types of labor—unskilled, semiskilled, and skilled—in the production of electronic components. The firm's production technology allows for the substitution of one type of labor for another. Also, the firm buys labor in a perfectly competitive resource market in which the price of labor stays the same regardless of the number of workers hired. The following table displays the marginal productivity and compensation in Canadian dollars for each type of labor.

What labor type should the firm hire when expanding output?

Type of Labor	Marginal Product (MP_{input}) per Day	Compensation (P_{input}) per Day ($)	$\frac{MP_{input}}{P_{input}}$
Unskilled (U)	200 units (MP_U)	100 (P_U)	2 units per $
Semiskilled (SS)	500 units (MP_{SS})	125 (P_{SS})	4 units per $
Skilled (S)	1,000 units (MP_S)	200 (P_S)	5 units per $

Solution: The firm minimizes cost and enhances profitability by adding skilled labor over the other two types because it has the highest ratio of *MP* to input price. As the marginal product of skilled labor declines with additional workers, MP_S/P_S decreases. When it declines to the same value as semiskilled labor, both skilled and semiskilled workers are added because their productivity per Canadian dollar of input cost is identical. Again, a diminishing marginal product decreases both ratios. When all three labor inputs have the same MP_{input}/P_{input}, the firm will add all three labor types at the same time when expanding output.

the input contributes to *TR*. It is also defined as the change in *TR* divided by the change in the quantity of the resource employed. If an input's *MRP* exceeds its cost, a contribution to profit is evident. For example, when the *MP* of the last unit of labor employed is 100 and the product price is 2.00, the *MRP* for that unit of labor (MRP_L) is 200. When the input price of labor is 125, the surplus value or contribution to profit is 75. In contrast, if *MRP* is less than the input's price, a loss would be incurred from employing that input unit. If the *MP* of the next unit of labor is 50 with a product price of 2.00, MRP_L will now be 100. With the same labor cost of 125, the firm would incur a loss of 25 when employing this input unit. Profit maximization occurs when the *MRP* equates to the price or cost of the input for each type of resource that is used in the production process.

In the case of multiple factor usage, the following equation holds true for n inputs:

$$\frac{MRP_1}{\text{Price of input 1}} = \ldots = \frac{MRP_n}{\text{Price of input } n} = 1 \tag{3-12}$$

When profit is maximized, *MRP* equals the input price for each type of resource used and all MRP_{input}/P_{input} are equal to 1.

EXAMPLE 3-12 Profit Maximization Using the Marginal Revenue Product and Resource Cost Approach

Using the data from the previous case of Canadian Global Electronic Corp., the following table shows the *MRP* per labor type when product price in Canadian dollars is \$0.50. *MRP* per day is calculated as the *MP* per type of labor from Example 3-11 multiplied by the product price.

Which type of labor contributes the most to profitability?

Type of Labor	Marginal Revenue Product (MRP_{input}) per Day (\$)	Compensation (P_{input}) per Day (\$)	$\frac{MRP_{input}}{P_{input}}$
Unskilled (*U*)	100 (MRP_U)	100 (P_U)	1.0
Semiskilled (*SS*)	250 (MRP_{SS})	125 (P_{SS})	2.0
Skilled (*S*)	500 (MRP_S)	200 (P_S)	2.5

Solution: Calculating the MRP_{input}/P_{input} values for the different labor categories yields ratio numbers of 1.0, 2.0, and 2.5 for unskilled, semiskilled, and skilled labor, respectively. The firm adds skilled labor first because it is the most profitable to employ, as indicated by MRP_S/P_S being the highest ratio of the three labor inputs. The contribution to profit by employing the next skilled worker is \$300, calculated as (\$500 − \$200). However, with the employment of additional skilled workers, MRP_S declines because of diminishing returns that are associated with the *MP* component. At the point where the skilled labor ratio drops below 2.0—for example, to 1.5—semiskilled labor becomes feasible to hire because its *MRP* exceeds its compensation by more than that of skilled labor.* Again, the diminishing returns effect decreases *MRP* when additional semiskilled workers are hired. In the case of unskilled labor, MRP_U equals the cost of labor; hence, no further contribution to profit accrues from adding this type of labor. In fact, adding another unskilled worker would probably reduce total profit because the next worker's compensation is likely to exceed *MRP* as a result of a declining *MP*. The input level that maximizes profit is where $MRP_U/P_U = MRP_{SS}/P_{SS} = MRP_S/P_S = 1$.

*The next semiskilled worker contributes \$125 (derived as \$250 − \$125) per day to profit, while the next skilled worker's contribution, based on a ratio of 1.5, is \$100 (MRP_S of \$300 minus compensation of \$200 per day).

4. SUMMARY

When assessing financial performance, a microeconomic exploration of a firm's profitability reveals more information to the analyst than does the typical macroeconomic examination of overall earnings. Crucial issues evolve when the firm fails to reward investors properly for their equity commitment and when the firm's operating status is not optimal in regard to resource employment, cost minimization, and profit maximization.

Among the points made in this chapter are the following:

- The two major concepts of profits are accounting profit and economic profit. Economic profit equals accounting profit minus implicit opportunity costs not included in accounting costs. Profit in the theory of the firm refers to economic profit.
- Normal profit is an economic profit of zero. A firm earning a normal profit is earning just enough to cover the explicit and implicit costs of resources used in running the firm, including, most importantly for publicly traded corporations, debt and equity capital.
- Economic profit is a residual value in excess of normal profit and results from access to positive NPV investment opportunities.
- The factors of production are the inputs to the production of goods and services and include land, labor, capital, and materials.
- Profit maximization occurs at the following points:
  - Where the difference between total revenue and total costs is the greatest.
  - Where marginal revenue equals marginal cost.
  - Where marginal revenue product equals the resource cost for each type of input.
- When total costs exceed total revenue, loss minimization occurs where the difference between total costs and total revenue is the least.
- In the long run, all inputs to the firm are variable, which expands profit potential and the number of cost structures available to the firm.
- Under perfect competition, long-run profit maximization occurs at the minimum point of the firm's long-run average total cost curve.
- In an economic loss situation, a firm can operate in the short run if total revenue covers variable cost but is inadequate to cover fixed cost; however, in the long run, the firm will exit the market if fixed costs are not covered in full.
- In an economic loss situation, a firm shuts down in the short run if total revenue does not cover variable cost in full, and eventually exits the market if the shortfall is not reversed.
- Economies of scale lead to lower average total cost; diseconomies of scale lead to higher average total cost.
- A firm's production function defines the relationship between total product and inputs.
- Average product and marginal product, which are derived from total product, are key measures of a firm's productivity.
- Increases in productivity reduce business costs and enhance profitability.
- An industry supply curve that is positively sloped in the long run will increase production costs to the firm. An industry supply curve that is negatively sloped in the long run will decrease production costs to the firm.
- In the short run, assuming constant resource prices, increasing marginal returns reduce the marginal costs of production, and decreasing marginal returns increase the marginal costs of production.

PRACTICE PROBLEMS[6]

1. Normal profit is *best* described as:
   A. zero economic profit.
   B. total revenue minus all explicit costs.
   C. the sum of accounting profit plus economic profit.

[6]These practice problems were developed by Christopher Anderson, CFA (Lawrence, Kansas, USA).

2. A firm supplying a commodity product in the marketplace is *most likely* to receive economic rent if:
   A. demand increases for the commodity and supply is elastic.
   B. demand increases for the commodity and supply is inelastic.
   C. supply increases for the commodity and demand is inelastic.

3. Entrepreneurs are *most likely* to receive payment or compensation in the form of:
   A. rent.
   B. profit.
   C. wages.

4. The marketing director for a Swiss specialty equipment manufacturer estimates the firm can sell 200 units and earn total revenue of CHF500,000. However, if 250 units are sold, revenue will total CHF600,000. The marginal revenue per unit associated with marketing 250 units instead of 200 units is *closest* to:
   A. CHF2,000.
   B. CHF2,400.
   C. CHF2,500.

5. An agricultural firm operating in a perfectly competitive market supplies wheat to manufacturers of consumer food products and animal feeds. If the firm were able to expand its production and unit sales by 10 percent, the *most likely* result would be:
   A. a 10 percent increase in total revenue.
   B. a 10 percent increase in average revenue.
   C. an increase in total revenue of less than 10 percent.

6. An operator of a ski resort is considering offering price reductions on weekday ski passes. At the normal price of €50 per day, 300 customers are expected to buy passes each weekday. At a discounted price of €40 per day, 450 customers are expected to buy passes each weekday. The marginal revenue per customer earned from offering the discounted price is *closest* to:
   A. €20.
   B. €40.
   C. €50.

7. The marginal revenue per unit sold for a firm doing business under conditions of perfect competition will *most likely* be:
   A. equal to average revenue.
   B. less than average revenue.
   C. greater than average revenue.

The following information relates to Questions 8 through 10.

A firm's director of operations gathers the following information about the firm's cost structure at different levels of output:

Exhibit A

Quantity (*Q*)	Total Fixed Cost (*TFC*)	Total Variable Cost (*TVC*)
0	200	0
1	200	100
2	200	150
3	200	200
4	200	240
5	200	320

8. Refer to the data in Exhibit A. When quantity produced is equal to four units, the average fixed cost (*AFC*) is *closest* to:
   A. 50.
   B. 60.
   C. 110.

9. Refer to the data in Exhibit A. When the firm increases production from four to five units, the marginal cost (*MC*) is *closest* to:
   A. 40.
   B. 64.
   C. 80.

10. Refer to the data in Exhibit A. The level of unit production resulting in the lowest average total cost (*ATC*) is *closest* to:
    A. 3.
    B. 4.
    C. 5.

11. The short-term breakeven point of production for a firm operating under perfect competition will *most likely* occur when:
    A. price is equal to average total cost.
    B. marginal revenue is equal to marginal cost.
    C. marginal revenue is equal to average variable costs.

12. The short-term shutdown point of production for a firm operating under perfect competition will *most likely* occur when:
    A. price is equal to average total cost.
    B. marginal revenue is equal to marginal cost.
    C. marginal revenue is less than average variable costs.

13. When total revenue is greater than total variable costs but less than total costs, in the short term a firm will *most likely*:
    A. exit the market.
    B. stay in the market.
    C. shut down production.

14. A profit maximum is *least likely* to occur when:
    A. average total cost is minimized.
    B. marginal revenue equals marginal cost.
    C. the difference between total revenue and total cost is maximized.

15. A firm that increases its quantity produced without any change in per-unit cost is experiencing:
    A. economies of scale.
    B. diseconomies of scale.
    C. constant returns to scale.

16. A firm is operating beyond minimum efficient scale in a perfectly competitive industry. To maintain long-term viability, the *most likely* course of action for the firm is to:
    A. operate at the current level of production.
    B. increase its level of production to gain economies of scale.
    C. decrease its level of production to the minimum point on the long-run average total cost curve.

17. Under conditions of perfect competition, in the long run firms will *most likely* earn:
    A. normal profits.
    B. positive economic profits.
    C. negative economic profits.

18. A firm engages in the development and extraction of oil and gas, the supply of which is price inelastic. The *most likely* equilibrium response in the long run to an increase in the demand for petroleum is that oil prices:
    A. increase, and extraction costs per barrel fall.
    B. increase, and extraction costs per barrel rise.
    C. remain constant, and extraction costs per barrel remain constant.

19. A firm develops and markets consumer electronic devices in a perfectly competitive, decreasing-cost industry. The firm's products have grown in popularity. The *most likely* equilibrium response in the long run to rising demand for such devices is for selling prices to:
    A. fall and per-unit production costs to decrease.
    B. rise and per-unit production costs to decrease.
    C. remain constant and per-unit production costs to remain constant.

The following information relates to Questions 20 and 21.

The manager of a small manufacturing firm gathers the following information about the firm's labor utilization and production:

Exhibit B

Labor (*L*)	Total Product (*TP*)
0	0
1	150
2	320
3	510
4	660
5	800

20. Refer to the data in Exhibit B. The number of workers resulting in the highest level of average product of labor is *closest* to:
    A. 3.
    B. 4.
    C. 5.

21. Refer to the data in Exhibit B. The marginal product of labor demonstrates increasing returns for the firm if the number of workers is *closest* to but not more than:
    A. 2.
    B. 3.
    C. 4.

22. A firm experiencing an increase in the marginal product of labor employed would *most likely*:
    A. allow an increased number of workers to specialize and become more adept at their individual functions.
    B. find that an increase in workers cannot be efficiently matched by other inputs that are fixed, such as property, plant, and equipment.
    C. find that the supply of skilled workers is limited, and additional workers lack essential skills and aptitudes possessed by the current workforce.

23. For a manufacturing company to achieve the most efficient combination of labor and capital and therefore to minimize total costs for a desired level of output, it will *most likely* attempt to equalize the:
    A. average product of labor to the average product of capital.
    B. marginal product per unit of labor to the marginal product per unit of capital.
    C. marginal product obtained per dollar spent on labor to the marginal product per dollar spent on capital.

24. A firm will expand production by 200 units and must hire at least one additional worker. The marginal product per day for one additional unskilled worker is 100 units, and for one additional skilled worker it is 200 units. Wages per day are $200 for an unskilled worker and $450 for a skilled worker. The firm will *most likely* minimize costs at the higher level of production by hiring:
    A. one additional skilled worker.
    B. two additional unskilled workers.
    C. either a skilled worker or two unskilled workers.

25. A Mexican firm employs unskilled, semiskilled, and skilled labor in a cost-minimizing mix at its manufacturing plant. The marginal product of unskilled labor is considerably lower than semiskilled and skilled labor, but the equilibrium wage for unskilled labor is only 300 pesos per day. The government passes a law that mandates a minimum wage of 400 pesos per day. Equilibrium wages for semiskilled and skilled labor exceed this minimum wage and therefore are not affected by the new law. The firm will *most likely* respond to the imposition of the minimum wage law by:
    A. employing more unskilled workers at its plant.
    B. employing fewer unskilled workers at its plant.
    C. keeping the mix of unskilled, semiskilled, and skilled workers the same.

The following information relates to Questions 26 and 27.

A firm produces handcrafted wooden chairs, employing both skilled craftspersons and automated equipment in its plant. The selling price of a chair is €100. A craftsperson earns €900 per week and can produce 10 chairs per week. Automated equipment leased for €800 per week also can produce 10 chairs per week.

26. The marginal revenue product (per week) of hiring an additional craftsperson is *closest* to:
    A. €100.
    B. €900.
    C. €1,000.

27. The firm would like to increase weekly output by 50 chairs. The firm would *most likely* enhance profits by:
    A. hiring additional craftspersons.
    B. leasing additional automated equipment.
    C. leasing additional automated equipment and hiring additional craftspersons in equal proportion.

CHAPTER 4

THE FIRM AND MARKET STRUCTURES

Richard G. Fritz

Michele Gambera, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Describe the characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly.
- Explain the relationships among price, marginal revenue, marginal cost, economic profit, and the elasticity of demand under each market structure.
- Describe the firm's supply function under each market structure.
- Describe and determine the optimal price and output for firms under each market structure.
- Explain factors affecting long-run equilibrium under each market structure.
- Describe pricing strategy under each market structure.
- Describe the use and limitations of concentration measures in identifying the various forms of market structure.
- Identify the type of market structure a firm is operating within.

1. INTRODUCTION

The purpose of this chapter is to build an understanding of the importance of market structure. As different market structures result in different sets of choices facing a firm's decision makers, an understanding of market structure is a powerful tool in analyzing issues such as a firm's pricing of its products and, more broadly, its potential to increase profitability. In the long run, a firm's profitability will be determined by the forces associated with the

market structure within which it operates. In a highly competitive market, long-run profits will be driven down by the forces of competition. In less competitive markets, large profits are possible even in the long run; in the short run, any outcome is possible. Therefore, understanding the forces behind the market structure will aid the financial analyst in determining firms' short- and long-term prospects.

Section 2 introduces the analysis of market structures. The section addresses questions such as: What determines the degree of competition associated with each market structure? Given the degree of competition associated with each market structure, what decisions are left to the management team developing corporate strategy? How does a chosen pricing and output strategy evolve into specific decisions that affect the profitability of the firm? The answers to these questions are related to the forces of the market structure within which the firm operates.

Sections 3, 4, 5, and 6 analyze demand, supply, optimal price and output, and factors affecting long-run equilibrium for perfect competition, monopolistic competition, oligopoly, and pure monopoly, respectively.

Section 7 reviews techniques for identifying the various forms of market structure. For example, there are accepted measures of market concentration that are used by regulators of financial institutions to judge whether a planned merger or acquisition will harm the competitive nature of regional banking markets. Financial analysts should be able to identify the type of market structure a firm is operating within. Each different structure implies a different long-run sustainability of profits. A summary and practice problems conclude the chapter.

2. ANALYSIS OF MARKET STRUCTURES

Traditionally, economists classify a market into one of four structures: perfect competition, monopolistic competition, oligopoly, and monopoly. Section 2.1 explains that four-way classification in more detail. Section 2.2 completes the introduction by providing and explaining the major points to evaluate in determining the structure to which a market belongs.

2.1. Economists' Four Types of Structure

Economists define a market as a group of buyers and sellers that are aware of each other and are able to agree on a price for the exchange of goods and services. Although the Internet has extended a number of markets worldwide, certain markets are limited by geographic boundaries. For example, the Internet search engine Google operates in a worldwide market. In contrast, the market for premixed cement is limited to the area within which a truck can deliver the mushy mix from the plant to a construction site before the compound becomes useless. Thomas L. Friedman's international best seller *The World Is Flat*[1] challenges the concept of the geographic limitations of the market. If the service being provided by the seller can be digitized, its market expands worldwide. For example, a technician can scan your injury in a clinic in Switzerland. That radiographic image can be digitized and sent to a radiologist in India to be read. As a customer (i.e., patient), you may never know that part of the medical service provided to you was the result of a worldwide market.

[1]Friedman (2006).

Some markets are highly concentrated, with the majority of total sales coming from a small number of firms. For example, in the market for small consumer batteries, three firms controlled 87 percent of the U.S. market as of 2005 (Duracell 43 percent, Energizer 33 percent, and Rayovac 11 percent). Other markets are very fragmented, such as automobile repairs, where small independent shops often dominate and large chains may or may not exist. New products can lead to market concentration: It is estimated that the Apple iPod had a world market share of over 70 percent among MP3 players in 2009.

The Importance of Market Structure

Consider the evolution of television broadcasting. As the market environment for television broadcasting evolved, the market structure changed, resulting in a new set of challenges and choices. In the early days, there was only one choice: the free analogue channels that were broadcast over the airwaves. In most countries, there was only one channel, owned and run by the government. In the United States, some of the more populated markets were able to receive more channels because local channels were set up to cover a market with more potential viewers. By the 1970s, new technologies made it possible to broadcast by way of cable connectivity, and the choices offered to consumers began to expand rapidly. Cable television challenged the free broadcast channels by offering more choice and a better-quality picture. The innovation was expensive for consumers and profitable for the cable companies. By the 1990s, a new alternative began to challenge the existing broadcast and cable systems: satellite television. Satellite providers offered a further expanded set of choices, albeit at a higher price than the free broadcast and cable alternatives. In the early 2000s, satellite television providers lowered their pricing to compete directly with the cable providers. Today, cable program providers, satellite television providers, and terrestrial digital broadcasters that offer premium and pay-per-view channels compete for customers, who are increasingly finding content on the Internet.

This is a simple illustration of the importance of market structure. As the market for television broadcasting became increasingly competitive, managers had to make decisions regarding product packaging, pricing, advertising, and marketing in order to survive in the changing environment.

Market structure can be broken down into four distinct categories: perfect competition, monopolistic competition, oligopoly, and monopoly.

We start with the most competitive environment, **perfect competition**. Unlike some economic concepts, perfect competition is not merely an ideal based on assumptions. Perfect competition is a reality—for example, in several commodities markets, where sellers and buyers have a strictly homogeneous product and no single producer is large enough to influence market prices. Perfect competition's characteristics are well recognized and its long-run outcome unavoidable. Profits under the conditions of perfect competition are driven to the required rate of return paid by the entrepreneur to borrow capital from investors (so-called normal profit or rental cost of capital). This does not mean that all perfectly competitive

industries are doomed to extinction by a lack of profits. On the contrary, millions of businesses that do very well are living under the pressures of perfect competition.

Monopolistic competition is also highly competitive; however, it is considered a form of imperfect competition. Two economists, Edward H. Chamberlin (United States) and Joan Robinson (United Kingdom), identified this hybrid market and came up with the term because there are strong elements of competition in this market structure and also some monopoly-like conditions. The competitive characteristic is a notably large number of firms, while the monopoly aspect is the result of product differentiation. That is, if the seller can convince consumers that its product is uniquely different from other, similar products, then the seller can exercise some degree of pricing power over the market. A good example is the brand loyalty associated with soft drinks such as Coca-Cola. Many of Coca-Cola's customers believe that the beverage is truly different from and better than all other soft drinks. The same is true for fashion creations and cosmetics.

The **oligopoly** market structure is based on a relatively small number of firms supplying the market. The small number of firms in the market means that each firm must consider what retaliatory strategies the other firms will pursue when prices and production levels change. Consider the pricing behavior of commercial airline companies. Pricing strategies and route scheduling are based on the expected reaction of the other carriers in similar markets. For any given route—say, from Paris, France, to Chennai, India—only a few carriers are in competition. If one of the carriers changes its pricing package, others will likely retaliate. Understanding the market structure of oligopoly markets can help in identifying a logical pattern of strategic price changes for the competing firms.

Finally, the least competitive market structure is **monopoly**. In pure monopoly markets, there are no other good substitutes for the given product or service. There is a single seller, which, if allowed to operate without constraint, exercises considerable power over pricing and output decisions. In most market-based economies around the globe, pure monopolies are regulated by a governmental authority. The most common example of a regulated monopoly is the local electrical power provider. In most cases, the monopoly power provider is allowed to earn a normal return on its investment, and prices are set by the regulatory authority to allow that return.

2.2. Factors That Determine Market Structure

Five factors determine market structure:

1. The number and relative size of firms supplying the product.
2. The degree of product differentiation.
3. The power of the seller over pricing decisions.
4. The relative strength of the barriers to market entry and exit.
5. The degree of nonprice competition.

The number and relative size of firms in a market influence market structure. If there are many firms, the degree of competition increases. With fewer firms supplying a good or service, consumers are limited in their market choices. One extreme case is the monopoly market structure, with only one firm supplying a unique good or service. Another extreme is perfect competition, with many firms supplying a similar product. Finally, an example of relative size is the automobile industry, in which a small number of large international producers (e.g., Ford and Toyota) are the leaders in the global market, and a number of small companies either

have market power because they are niche players (e.g., Ferrari) or have little market power because of their narrow range of models or limited geographical presence (e.g., Škoda).

In the case of monopolistic competition, there are many firms providing products to the market, as with perfect competition. However, one firm's product is differentiated in some way that makes it appear better than similar products from other firms. If a firm is successful in differentiating its product, the differentiation will provide pricing leverage. The more dissimilar the product appears, the more the market will resemble the monopoly market structure. A firm can differentiate its product through aggressive advertising campaigns, frequent styling changes, the linking of its product with complementary products, or a host of other methods.

When the market dictates the price based on aggregate supply and demand conditions, the individual firm has no control over pricing. The typical hog farmer in Nebraska and the milk producer in Bavaria are **price takers**. That is, they must accept whatever price the market dictates. This is the case under the market structure of perfect competition. In the case of monopolistic competition, the success of product differentiation determines the degree with which the firm can influence price. In the case of oligopoly, there are so few firms in the market that price control becomes possible. However, the small number of firms in an oligopoly market invites complex pricing strategies. Collusion, price leadership by dominant firms, and other pricing strategies can result.

The degree to which one market structure can evolve into another and the difference between potential short-run outcomes and long-run equilibrium conditions depend on the strength of the barriers to entry and the possibility that firms fail to recoup their original costs or lose money for an extended period of time and are therefore forced to exit the market. Barriers to entry can result from very large capital investment requirements, as in the case of petroleum refining. Barriers may also result from patents, as in the case of some electronic products and prescription drug formulas. Another entry consideration is the possibility of high exit costs. For example, plants that are specific to a special line of products, such as aluminum smelting plants, are nonredeployable, and exit costs would be high without a liquid market for the firm's assets. High exit costs deter entry and are therefore also considered barriers to entry. In the case of farming, the barriers to entry are low. Production of corn, soybeans, wheat, tomatoes, and other produce is an easy process to replicate; therefore, those are highly competitive markets.

Nonprice competition dominates those market structures where product differentiation is critical. Therefore, monopolistic competition relies on competitive strategies that may not include pricing changes. An example of nonprice competition is product differentiation through marketing. In other circumstances, nonprice competition may occur because the few firms in the market feel dependent on each other. Each firm fears retaliatory price changes that would reduce total revenue for all of the firms in the market. Because oligopoly industries have so few firms, each firm feels dependent on the pricing strategies of the others. Therefore, nonprice competition becomes a dominant strategy.

From the perspective of the owners of the firm, the most desirable market structure is that with the most control over price, because this control can lead to large profits. Monopoly and oligopoly markets offer the greatest potential control over price; monopolistic competition offers less control. Firms operating under perfectly competitive market conditions have no control over price. From the consumers' perspective, the most desirable market structure is that with the greatest degree of competition, because prices are generally lower. Thus, consumers would prefer as many goods and services as possible to be offered in competitive markets.

As often happens in economics, there is a trade-off. While perfect competition gives the largest quantity of a good at the lowest price, other market forms may spur more innovation. Specifically, there may be high costs in researching a new product, and firms will incur such costs only if they expect to earn an attractive return on their research investment. This is the case often made for medical innovations, for example—the cost of clinical trials and experiments to create new medicines would bankrupt perfectly competitive firms but may be acceptable in an oligopoly market structure. Therefore, consumers can benefit from less than perfectly competitive markets.

Porter's Five Forces and Market Structure

A financial analyst aiming to establish market conditions and consequent profitability of incumbent firms should start with the questions framed by Exhibit 4-1: how many sellers there are, whether the product is differentiated, and so on. Moreover, in the case of monopolies and quasi-monopolies, the analyst should evaluate the legislative and regulatory framework: Can the company set prices freely, or are there governmental controls? Finally, the analyst should consider the threat of competition from potential entrants.

EXHIBIT 4-1 Characteristics of Market Structure

Market Structure	Number of Sellers	Degree of Product Differentiation	Barriers to Entry	Pricing Power of Firm	Nonprice Competition
Perfect competition	Many	Homogeneous/ standardized	Very low	None	None
Monopolistic competition	Many	Differentiated	Low	Some	Advertising and product differentiation
Oligopoly	Few	Homogeneous/ standardized	High	Some or considerable	Advertising and product differentiation
Monopoly	One	Unique product	Very high	Considerable	Advertising

This analysis is often summarized by students of corporate strategy as Porter's five forces, named after Harvard Business School professor Michael E. Porter. His book *Competitive Strategy* (1980) presented a systematic analysis of the practice of market strategy. The five forces are:

1. Threat of substitutes.
2. Threat of entry.
3. Intensity of competition among incumbents.
4. Bargaining power of customers.
5. Bargaining power of suppliers.

It is easy to note the parallels between four of these five forces and the columns in Exhibit 4-1. The only "orphan" is the power of suppliers, which is not at the core of the theoretical economic analysis of competition, but has substantial weight in the practical analysis of competition and profitability.

Some stock analysts (e.g., Dorsey 2004) use the term *economic moat* to suggest that there are factors protecting the profitability of a firm that are similar to the ditches full of water that used to provide protection for some medieval castles. A deep moat means that there is little or no threat of entry by invaders (i.e., competitors). It also means that customers are locked in because of high switching costs.

3. PERFECT COMPETITION

Perfect competition is characterized by the five conditions presented in Exhibit 4-1:

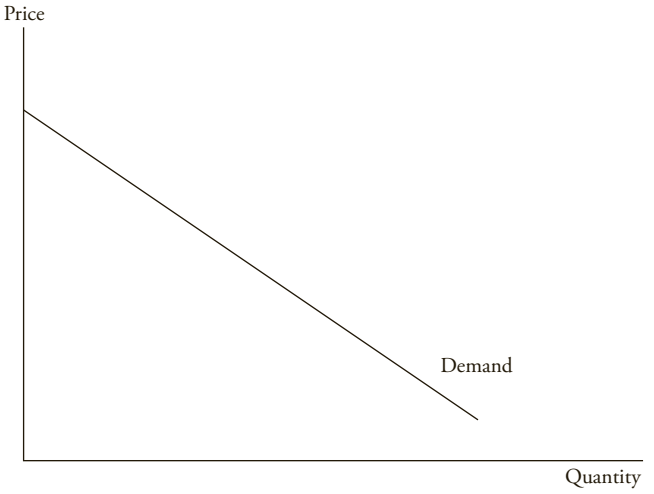
1. There is a large number of potential buyers and sellers.
2. The products offered by the sellers are virtually identical.
3. There are few or easily surmountable barriers to entry and exit.
4. Sellers have no market-pricing power.
5. Nonprice competition is absent.

While few markets achieve the distinction of being perfectly competitive, it is useful to establish the outcome associated with this market structure as a benchmark against which other market structures can be compared. The most typical example of perfect competition is found in certain aspects of the agriculture industry, such as the large number of farmers growing corn for animal feed. Corn is a primary source of food for pork, beef, and poultry production. A bushel of corn from Farmer Brown is virtually identical to a bushel of corn from Farmer Lopez. If hog farmers need corn to feed their hogs, it does not matter whether the corn comes from Farmer Brown or from Farmer Lopez. Furthermore, the aggregate corn market is well defined, with active futures and spot markets. Information about the corn market is easy and inexpensive to access, and there is no way to differentiate the product, such as by advertising. Agribusiness is capital intensive, but where arable land is relatively abundant and water is available, the barriers to entry (e.g., capital and expertise) for corn production are relatively low.

3.1. Demand Analysis in Perfectly Competitive Markets

The price of a homogeneous product sold in a competitive market is determined by the demand and supply in that market. Economists usually represent demand and supply in a market through demand and supply curves in a two-axis plane, where quantity and price are shown on the *x*-axis and *y*-axis, respectively. Economists believe that demand functions have negative slopes, as shown in Exhibit 4-2. That is, at high prices, less is demanded. For normal goods and services, as the price declines, the quantity demanded increases. This concept is based on two effects: the income effect and the substitution effect. The income effect results from the increased purchasing power the consumer has when prices fall. With lower prices, the

EXHIBIT 4-2 Market Demand in Perfect Competition



consumer can afford to purchase more of the product. The substitution effect comes from the increasing attractiveness of the lower-priced product. If soybean prices are unchanged and corn prices decrease, hog farmers will substitute corn for soybeans as feed for their animals.

Assume the demand for this product can be specified as:

$$Q_D = 50 - 2P$$

where Q_D is the quantity of demand and P is the product's price. This demand function can be rearranged in terms of price:

$$P = 25 - 0.5Q_D$$

In this form, total revenue (TR) is equal to price times quantity, or $P \times Q_D$. Thus,

$$\text{TR} = PQ_D = 25Q_D - 0.5Q_D^2$$

Average revenue (AR) can be found by dividing TR by Q_D. Therefore,

$$\text{AR} = \text{TR}/Q_D = (25Q_D - 0.5Q_D^2)/Q_D = 25 - 0.5Q_D$$

Note that the AR function is identical to the market demand function. The assumption here is that the relationship between price and quantity demanded is linear. Clearly, that may not be the case in the real market. Another simplifying assumption made is that the price of the product is the only determinant of demand. Again, that is not likely in the real market. For example, economic theory suggests that consumer income is another important factor in determining demand. The prices of related goods and services, such as substitutes and complements, are also considered factors affecting demand for a particular product.

Marginal revenue (MR) is the change in total revenue per extra increment sold when the quantity sold changes by a small increment, ΔQ_D. Substituting $(Q_D + \Delta Q_D)$ into the total revenue (TR) equation, marginal revenue can be expressed as:

$$\text{MR} = \frac{\Delta \text{TR}}{\Delta Q_D} = \frac{[25(Q_D + \Delta Q_D) - 0.5(Q_D^2 + 2Q_D\Delta Q_D + \Delta Q_D^2)] - [25Q_D - 0.5Q_D^2]}{\Delta Q_D}$$

$$= \frac{25\Delta Q_D - Q_D\Delta Q_D - 0.5\Delta Q_D^2}{\Delta Q_D} = 25 - Q_D - 0.5\Delta Q_D$$

For example, suppose $Q_D = 5$ and $\Delta Q_D = 1$; then total revenue increases from 112.50 $[= 25(5) - 0.5\ (5^2)]$ to 132 $[= 25(6) - 0.5(6^2)]$, and marginal revenue is $19.5 = (132 - 112.5)/1$. Note that marginal revenue is equal to $(25 - Q_D - 0.5\Delta Q_D)$. Now suppose that ΔQ_D is much smaller, for example $\Delta Q_D = 0.1$. In this case, total revenue increases to 114.495 $[= 25(5.1) - 0.5(5.1^2)]$, and marginal revenue is $1.995/0.1 = 19.95$. It is straightforward to confirm that as ΔQ_D gets smaller marginal revenue gets closer to $20 = 25 - Q_D$. So, for very small changes in the quantity sold we can write marginal revenue as:[2]

$$\text{MR} = 25 - Q_D$$

Although we have introduced the concept of marginal revenue in the context of the demand curve for the market as a whole, its usefulness derives from its role in the output and pricing decisions of individual firms. As we will see, marginal revenue and an analogous concept, marginal cost, are critical in determining firms' profit-maximizing strategies.

3.1.1. Elasticity of Demand

Consumers respond differently to changes in the prices of different kinds of products and services. The quantity demanded for some products is very price sensitive, while for other products, price changes result in little change in the quantity demanded. Economists refer to the relationship between changes in price and changes in the quantity demanded as the price elasticity of demand. Therefore, the demand for the former group of products—those that are very price sensitive—is said to have high price elasticity, whereas the demand for the latter group is said to have low price elasticity. Understanding the sensitivity of demand changes to changes in price is critical to understanding market structures.

Price elasticity of demand measures the percentage change in the quantity demanded given a percentage change in the price of a given product. Because the relationship of demand to price is negative, the price elasticity of demand would be negative. *Many economists, however, present the price elasticity as an absolute value, so that price elasticity has a positive sign. We will follow that convention.* Higher price elasticity indicates that consumers are very responsive to changes in price. Lower values for price elasticity imply that consumers are not very responsive to price changes. Price elasticity can be measured with the following relationship:

$$\varepsilon_P = -(\%\text{ change in } Q_D) \div (\%\text{ change in } P)$$

[2]Readers who are familiar with calculus will recognize this as the derivative of total revenue with respect to the quantity sold.

where ε_P is price elasticity of demand, Q_D is the quantity demanded, and P is the product's price.

Price elasticity of demand falls into three categories. When demand is very responsive to price change, it is identified as *elastic*. When demand is not responsive to price change, it is identified as *inelastic*. When the percentage change in quantity demanded is exactly the same as the percentage change in price, the demand is called *unitary elastic*.

1. If $\varepsilon_P > 1$, demand is elastic.
2. If $\varepsilon_P = 1$, demand is unitary elastic.
3. If $\varepsilon_P < 1$, demand is inelastic.

Price elasticity of demand depends on several factors. *Price elasticity will be higher if there are many close substitutes for the product*. If a product has many good alternatives, consumers will be more sensitive to price changes. For example, carbonated beverages (soft drinks) have many close substitutes. It takes strong brand loyalty to keep customer demand high in the soft drink market when one brand's price is strategically lowered; the price elasticity of demand for Coca-Cola has been estimated to be 3.8. For products with numerous close substitutes, demand is highly elastic. For products with few close substitutes, demand is lower in price elasticity and would be considered price inelastic. The demand for first-class airline tickets is often seen as inelastic because only very wealthy people are expected to buy them; the demand for economy-class tickets is elastic because the typical consumer for this product is more budget-conscious. Consumers do not consider economy-class airline tickets a close substitute for first-class accommodations, particularly on long flights.

The airline ticket example introduces another determinant of price elasticity of demand. *The greater the share of the consumer's budget spent on the item, the higher the price elasticity of demand*. Expensive items, such as durable goods (e.g., refrigerators and televisions), tend to have higher elasticity measures, while less expensive items, such as potatoes and salt, have lower elasticity values. Consumers will not change their normal salt consumption if the price of salt decreases by 10 percent. Instead, they will buy their next package of salt when they are about to run out, with very little regard to the price change.

The airline ticket also makes a good example for the final factor determining price elasticity. *Price elasticity of demand also depends on the length of time within which the demand schedule is being considered*. Holiday airline travel is highly price elastic. Consumers shop vigorously for vacation flights because they have time to plan their holidays. Business airline travelers typically have less flexibility in determining their schedules. If your business requires a face-to-face meeting with a client, then the price of the ticket is somewhat irrelevant. If gasoline prices increase, there is very little you can do in the short run but pay the higher price. However, evidence of commuter choices indicates that many use alternative transportation methods after the gasoline price spikes. In the long run, higher gasoline prices will lead consumers to change their modes of transportation, trading in less efficient vehicles for automobiles with higher-rated gas mileage or public transit options where available.

There are two extreme cases of price elasticity of demand. One extreme is the **horizontal demand schedule**. This term implies that at a given price, the response in the quantity demanded is infinite. *This is the demand schedule faced by a perfectly competitive firm, because it is a price taker*, as in the case of a corn farmer. If the corn farmer tried to charge a higher price than the market price, nobody would buy her product. On the other hand, the farmer has no incentive to sell at a lower price because she can sell all she can produce at the market price. In

EXHIBIT 4-3 Empirical Price Elasticities

Commodity (Good/Service)	Price Elasticity of Market Demand
Alcoholic beverages consumed at home	
Beer	0.84
Wine	0.55
Liquor	0.50
Coffee	
Regular	0.16
Instant	0.36
Credit charges on bank cards	2.44
Furniture	3.04
Glassware/china	1.20
International air transportation, United States/Europe	1.20
Shoes	0.73
Soybean meal	1.65
Tomatoes	2.22

Various sources, as noted in McGuigan, Moyer, and Harris (2008, 95). These are the elasticities with respect to the product's own price; by convention, they are shown here as positive numbers.

a perfectly competitive market the quantity supplied by an individual firm has a negligible effect on the market price. In the case of *perfect price elasticity*, the measure is $\varepsilon_P = \infty$.

The other extreme is the **vertical demand schedule**. The vertical demand schedule implies that some fixed quantity is demanded, regardless of price. An example of such demand is the diabetic consumer with the need for a certain amount of insulin. If the price of insulin goes up, the patient will not consume less of it. The amount desired is set by the patient's medical condition. The measure for *perfect price inelasticity* is $\varepsilon_P = 0$.

The nature of the elasticity calculation and consumer behavior in the marketplace imply that for virtually any product (excluding cases of perfect elasticity and perfect inelasticity) demand is more elastic at higher prices and less elastic (more inelastic) at lower prices. For example, at current low prices, the demand for table salt is very inelastic. However, if table salt increased in price to hundreds of dollars per ounce, consumers would become more responsive to its price changes. Exhibit 4-3 reports several empirical estimates of price elasticity of demand.

3.1.2. Other Factors Affecting Demand

There are two other important forces that influence shifts in consumer demand. One influential factor is consumer income and the other is the price of a related product. For normal goods, as consumer income increases, the demand increases. The degree to which consumers respond to higher incomes by increasing their demand for goods and services is referred to as income elasticity of demand. **Income elasticity of demand** measures the

responsiveness of demand to changes in income. The calculation is similar to that of price elasticity, with the percentage change in income replacing the percentage change in price. Note the new calculation:

$$\varepsilon_Y = (\%\text{ change in } Q_D) \div (\%\text{ change in } Y)$$

where ε_Y is income elasticity of demand, Q_D is the quantity demanded, and Y is consumer income. For normal goods, the measure ε_Y will be a positive value. That is, as consumers' income rises, more of the product is demanded. For products that are considered luxury items, the measure of income elasticity will be greater than 1. There are other goods and services that are considered inferior products. For inferior products, as consumer income rises, less of the product is demanded. Inferior products will have negative values for income elasticity. For example, a person on a small income may watch television shows, but if this person had more income, she would prefer going to live concerts and theater performances; in this example, television shows would be the inferior good.

As a technical issue, the difference between price elasticity of demand and income elasticity of demand is that the demand adjustment for price elasticity represents a movement *along the demand schedule* because the demand schedule represents combinations of price and quantity. The demand adjustment for income elasticity represents a *shift in the demand curve* because with a higher income one can afford to purchase more of the good at any price. For a normal good, an increase in income would shift the demand schedule out to the right, away from the origin of the graph, and a decrease in income would shift the demand curve to the left, toward the origin.

The final factor influencing demand for a product is the change in price of a related product, such as a strong substitute or a complementary product. If a close competitor in the beverage market lowers its price, some consumers may substitute that product for your product. Thus, your product's demand curve will shift to the left, toward the origin of the graph. **Cross-price elasticity of demand** is the responsiveness of the demand for product A that is associated with the change in price of product B:

$$\varepsilon_X = (\%\text{ change in } Q_{DA}) \div (\%\text{ change in } P_B)$$

where ε_X is cross-price elasticity of demand, Q_{DA} is the quantity demanded of product A, and P_B is the price of product B.

When the cross-price elasticity of demand between two products is *positive*, the two products are considered to be **substitutes**. For example, you may expect to have positive cross-price elasticity between honey and sugar. If the measure of cross-price elasticity is *negative*, the two products are referred to as **complements** of each other. For example, if the price of DVDs goes up, you would expect consumers to buy fewer DVD players. In this case, the cross-price elasticity of demand would have a negative value.

Reviewing cross-price elasticity values provides a simple test for the degree of competition in the market. The more numerous and the closer the substitutes for a product, the lower the pricing power of firms selling in that market; the fewer the substitutes for a product, the greater the pricing power. One interesting application was a U.S. Supreme Court case involving the production and sale of cellophane by DuPont.[3] The court noted that the

[3] *U.S. v. DuPont*, 351 U.S. 377 (1956), as noted in McGuigan, Moyer, and Harris (2008).

relevant product market for DuPont's cellophane was the broader flexible packaging materials market. The Supreme Court found the cross-price elasticity of demand between cellophane and other flexible packaging materials to be sufficiently high and exonerated DuPont from a charge of monopolizing the market.

Because price elasticity of demand relates changes in price to changes in the quantity demanded, there must be a logical relationship between marginal revenue and price elasticity. Recall that marginal revenue equals the change in total revenue given a change in output or sales. An increase in total revenue results from a decrease in price that results in an increase in sales. In order for the increase in the quantity demanded to be sufficient to offset the decline in price, the percentage change in quantity demanded must be greater than the percentage decrease in price. The relationship between TR and price elasticity is as follows:

$$\begin{aligned}
&\varepsilon_P > 1 \text{ Demand is elastic} && \uparrow P \rightarrow \text{TR} \downarrow \text{ and } \downarrow P \rightarrow \text{TR} \uparrow \\
&\varepsilon_P = 1 \text{ Demand is unitary elastic} && \updownarrow P \rightarrow \text{no change in TR} \\
&\varepsilon_P < 1 \text{ Demand is inelastic} && \uparrow P \rightarrow \text{TR} \uparrow \text{ and } \downarrow P \rightarrow \text{TR} \downarrow
\end{aligned}$$

Total revenue is maximized when marginal revenue is zero. The logic is that as long as marginal revenue is positive (i.e., each additional unit sold contributes to additional total revenue), total revenue will continue to increase. Only when marginal revenue becomes negative will total revenue begin to decline. Therefore, the percentage decrease in price is greater than the percentage increase in quantity demanded. The relationship between MR and price elasticity can be expressed as:

$$\text{MR} = P[1 - (1/\varepsilon_P)]$$

An understanding of price elasticity of demand is an important strategic tool. It would be very useful to know in advance what would happen to your firm's total revenue if you increased the product's price. If you are operating in the inelastic portion of the demand curve, increasing the price of the product will increase total revenue. However, if you are operating in the elastic portion of the product's demand curve, increasing the price will decrease total revenue.

Decision makers can also use the relationship between marginal revenue and price elasticity of demand in other ways. For example, suppose you are a farmer considering planting soybeans or some other feed crop, such as corn. From Exhibit 4-3, we know that soybean meal's price elasticity of demand has been estimated to be 1.65. We also know that the current (August 2010) soybean meal price is $330.14 per metric ton. Therefore, by solving the preceding equation, we find that the expected marginal revenue per metric ton of soybean meal is $130.05. Soybeans may prove to be a profitable crop for the farmer. However, just a few years earlier, in August 2006, the price of a metric ton of soybean meal was $175.91. Given the crop's price elasticity of demand, the estimated marginal revenue per metric ton was then $69.30. The lower price translates into lower marginal revenue and might induce the farmer to plant a more profitable feed crop instead.

How do business decision makers decide what level of output to bring to the market? To answer that question, the firm must understand its cost of resources, its production relations, and its supply function. Once the supply function is well defined and understood, it is combined with the demand analysis to determine the profit-maximizing levels of output.

3.1.3. Consumer Surplus: Value minus Expenditure

To this point, we have discussed the fundamentals of supply and demand curves and explained a simple model of how a market can be expected to arrive at an equilibrium combination of price and quantity. While it is certainly necessary for the analyst to understand the basic workings of the market model, it is also crucial to have a sense of why we might care about the nature of the equilibrium. In this section we review the concept of **consumer surplus**, which is helpful in understanding and evaluating business pricing strategies. Consumer surplus is defined as the difference between the value that a consumer places on the units purchased and the amount of money that was required to pay for them. It is a measure of the value gained by the buyer from the transaction.

To get an intuitive feel for the concept of consumer surplus, consider the last thing you purchased. Whatever it was, think of how much you actually paid for it. Now contrast that price with the maximum amount you *would have been willing to pay* rather than go without the item altogether. If those two numbers are different, we say you received some consumer surplus from your purchase. You got a bargain because you would have been willing to pay more than you had to pay.

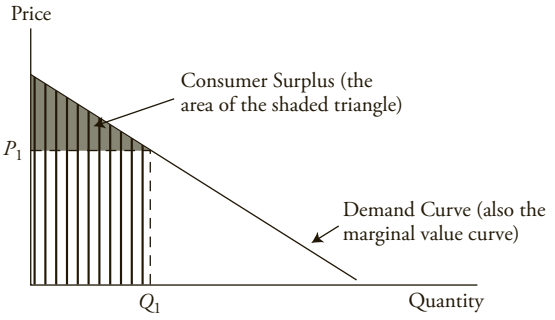
Earlier, we referred to the law of demand, which says that as the price falls, consumers are willing to buy more of the good. This observation translates into a negatively sloped demand curve. Alternatively, we could say that the highest price that consumers are willing to pay for an additional unit declines as they consume more and more of a good. In this way, we can interpret their *willingness to pay* as a measure of how much they *value* each additional unit of the good. This is a very important point: In order to purchase a unit of some good, consumers must give up something else they value. So the price they are willing to pay for an additional unit of a good is a measure of how much they value that unit, in terms of the other goods they must sacrifice to consume it.

If demand curves are negatively sloped, it must be because the value of each additional unit of the good falls as more of the good is consumed. We will explore this concept further, but for now it is enough to recognize that the demand curve can therefore be considered a **marginal value curve**, because it shows the highest price consumers would be willing to pay for each additional unit. In effect, the demand curve is the willingness of consumers to pay for each additional unit.

This interpretation of the demand curve allows us to measure the total value of consuming any given quantity of a good: It is the sum of all the marginal values of each unit consumed, up to and including the last unit. Graphically, this measure translates into the area under the consumer's demand curve, up to and including the last unit consumed, as shown in Exhibit 4-4, where the consumer is choosing to buy Q_1 units of the good at a price of P_1. The marginal value of the Q_1th unit is clearly P_1, because that is the highest price the consumer is willing to pay for that unit. Importantly, however, the marginal value of each unit *up to* the Q_1th is greater than P_1.

Because the consumer would have been willing to pay more for each of those units than she actually paid (P_1), we can say she received more value than the cost to her of buying them. This extra value is the buyer's consumer surplus. The *total value* of quantity Q_1 to the buyer is the area of the vertically lined trapezoid in Exhibit 4-4. The *total expenditure* is only the area of the rectangle with height P_1 and base Q_1 (bottom section). The total consumer surplus received from buying Q_1 units at a level price of P_1 per unit is the difference between the area under the demand curve and the area of the rectangle $P_1 \times Q_1$. The resulting area is shown as the shaded triangle (upper section).

EXHIBIT 4-4 Consumer Surplus

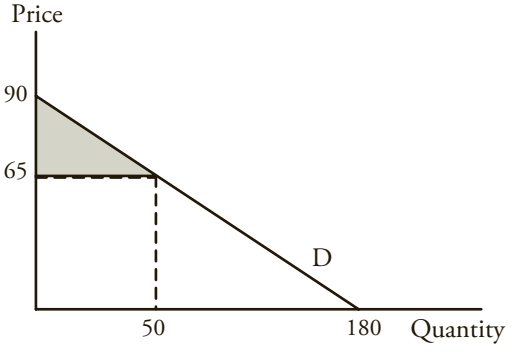


Note: Consumer surplus is the area beneath the demand curve and above the price paid.

EXAMPLE 4-1 Consumer Surplus

A market demand function is given by the equation $Q_D = 180 - 2P$. Find the value of consumer surplus if price is equal to 65.

Solution: First, input 65 into the demand function to find the quantity demanded at that price: $Q_D = 180 - 2(65) = 50$. Then, to make drawing the demand curve easier, invert the demand function by solving for P in terms of Q_D: $P = 90 - 0.5Q_D$. Note that the price intercept is 90 and the quantity intercept is 180. Draw the demand curve:



Find the area of the triangle above the price of 65 and below the demand curve, up to quantity 50: Area = ½(Base)(Height) = ½(50)(25) = 625.

3.2. Supply Analysis in Perfectly Competitive Markets

Consider two corn farmers, Mr. Brown and Ms. Lopez. They both have land available to them to grow corn and can sell at one price, say 3 currency units per kilogram. They will try to produce as much corn as is profitable at that price. If the price is driven up to 5 currency units per kilogram by new consumers entering the market—say, ethanol producers—Mr. Brown and Ms. Lopez will try to produce more corn. To increase their output levels, they may have to use less productive land, increase irrigation, use more fertilizer, or all three. Their production costs will likely increase. They will both still try to produce as much corn as possible in order to profit at the new, higher price of 5 currency units per kilogram. Exhibit 4-5 illustrates this example. Note that the supply functions for the individual firms have positive slopes. Thus, as prices increase, the firms supply greater quantities of the product.

Notice that the market supply curve is the sum of the supply curves of the individual firms—Brown, Lopez, and others—that make up the market. Assume that the supply function for the market can be expressed as a linear relationship, as follows:

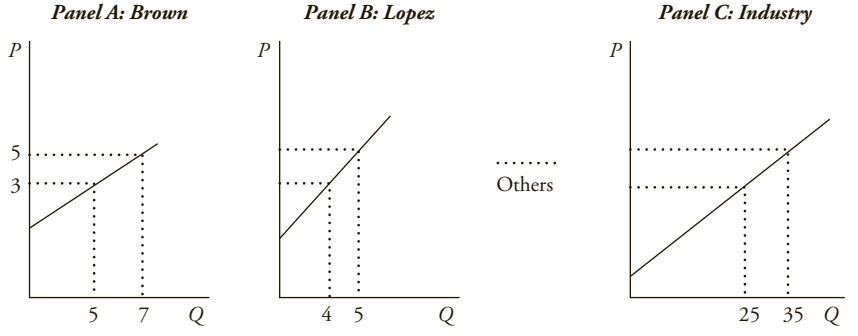
$$Q_S = 10 + 5P, \text{ or } P = -2 + 0.2Q_S$$

where Q_S is the quantity supplied and P is the price of the product.

Before we analyze the optimal supply level for the firm, we need to point out that economic costs and profits differ from accounting costs and profits in a significant way. **Economic costs** include all the remuneration needed to keep the productive resource in its current employment or to acquire the resource for productive use.

In order to evaluate the remuneration needed to keep the resource in its current use and attract new resources for productive use, economists refer to the resource's **opportunity cost**. Opportunity cost is measured by determining the resource's next best opportunity. If a corn farmer could be employed in an alternative position in the labor market with an income of 50,000, then the opportunity cost of the farmer's labor is 50,000. Similarly, the farmer's land and capital could be leased to another farmer or sold and reinvested in another type of business. The return forgone by not doing so is an opportunity cost. In economic terms, total cost includes the full normal market return on all the resources utilized in the business.

EXHIBIT 4-5 Firm and Market Supply in Perfect Competition



Economic profit is the difference between total revenue (TR) and total cost (TC). Economic profit differs from accounting profit because accounting profit does not include opportunity cost. Accounting profit includes only explicit payments to outside providers of resources (e.g., workers, vendors, lenders) and depreciation based on the historical cost of physical capital.

3.3. Optimal Price and Output in Perfectly Competitive Markets

Carrying forward our examples from Sections 3.1 and 3.2, we can now combine the market supply and demand functions to solve for the equilibrium price and quantity, where Q^* represents the equilibrium level of both supply and demand.

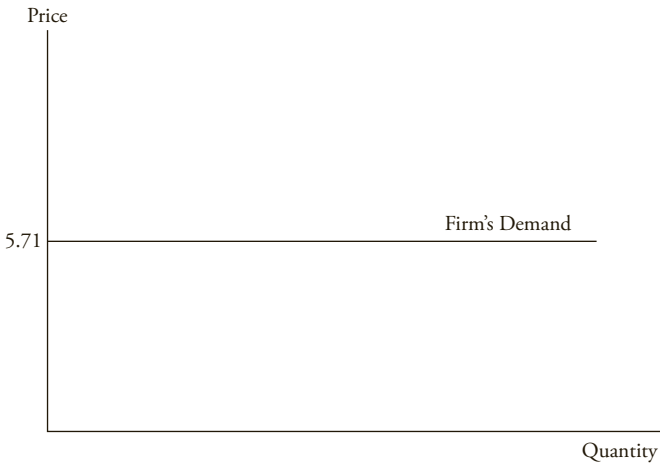
$$\begin{aligned} P = 25 - 0.5Q_D &= -2 + 0.2Q_S = P \\ 25 - 0.5Q_D &= -2 + 0.2Q_S \\ 27 &= 0.7Q^* \\ Q^* &= 38.57 \end{aligned}$$

According to the market demand curve, the equilibrium price is:

$$P = 25 - 0.5Q^* = 25 - 0.5(38.57) = 25 - 19.29 = 5.71$$

With many firms in the market and total output in the market of almost 39 units of the product, the effective market price would be 5.71. This result becomes the demand function for each perfectly competitive firm. Even if a few individual producers could expand production, there would not be a noticeable change in the market equilibrium price. In fact, if any one firm could change the equilibrium market price, the market would not be in perfect competition. Therefore, the demand curve that each perfectly competitive firm faces is a horizontal line at the equilibrium price, as shown in Exhibit 4-6, even though the demand curve for the whole market is downward sloping.

EXHIBIT 4-6 Individual Firm's Demand in Perfect Competition



EXAMPLE 4-2 Demand Curves in Perfect Competition

Is it possible that the demand schedule faced by Firm A is horizontal while the demand schedule faced by the market as a whole is downward sloping?

A. No, because Firm A can change its output based on demand changes.
B. No, because a horizontal demand curve means that elasticity is infinite.
C. Yes, because consumers can go to another firm if Firm A charges a higher price, and Firm A can sell all it produces at the market price.

Solution: C is correct. Firm A cannot successfully charge a higher price and has no incentive to sell at a price below the market price.

To analyze the firm's revenue position, recall that average revenue (AR) is equivalent to the firm's demand function. Therefore, the horizontal line that represents the firm's demand curve is the firm's AR schedule.

Marginal revenue (MR) is the incremental increase in total revenue (TR) associated with each additional unit sold. For every extra unit the firm sells, it receives 5.71. Thus, the firm's MR schedule is also the horizontal line at 5.71. TR is calculated by multiplying AR by the quantity of products sold. Total revenue is the area under the AR line at the point where the firm produces the output. In the case of perfect competition, the following conditions hold for the individual firm:

$$\text{Price} = \text{Average revenue} = \text{Marginal revenue}$$

The next step is to develop the firm's cost functions. The firm knows that it can sell the entire product it produces at the market's equilibrium price. How much should it produce? That decision is determined by analysis of the firm's costs and revenues. A corn farmer uses three primary resources: land, labor, and capital. In economics, capital is any man-made aid to production. The corn farmer's capital includes the irrigation system, tractors, harvesters, trucks, grain bins, fertilizer, and so forth. The labor includes the farmer, perhaps members of the farmer's family, and hired labor. In the initial stages of production, only the farmer and the farmer's family are cultivating the land, with a significant investment in capital. They have a tractor, fertilizer, irrigation equipment, grain bins, seed, and a harvester. The investment in land and capital is relatively high compared with the labor input. In this production phase, the average cost of producing a bushel of corn is high. As they begin to expand by adding labor to the collection of expensive land and capital, the average cost of producing corn begins to decline—for example, because one tractor can be used more intensively to plow a larger amount of land. When the combination of land, labor, and capital approaches an efficient range, the average cost of producing a bushel of corn declines.

Given a certain level of technology, there is a limit to the increase in productivity. Eventually something begins to cause declining marginal productivity. That is, each additional unit of input produces a progressively smaller increase in output. This force is called the **law of diminishing returns**. This law helps define the shape of the firm's cost functions. Average cost

and marginal cost will be U-shaped. Over the initial stages of output, average and marginal costs will decline. At some level of output, the law of diminishing returns will overtake the efficiencies in production, and average and marginal costs will increase.

Average cost (AC) is total cost (TC) divided by output (Q). Therefore,

$$\text{AC} = \text{TC}/Q$$

Note that we have defined average cost (AC) in terms of total costs. Many authors refer to this as "average total cost" to distinguish it from a related concept, average variable cost, which omits fixed costs. In the remainder of this chapter, *average cost* should be understood to mean *average total cost*.

Marginal cost (MC) is the change in TC associated with an incremental change in output:

$$\text{MC} = \Delta\text{TC}/\Delta Q$$

By definition, fixed costs do not vary with output, so marginal cost reflects only changes in variable costs.[4] MC declines initially because processes can be made more efficient and specialization makes workers more proficient at their tasks. However, at some higher level of output, MC begins to increase (e.g., must pay workers a higher wage to have them work overtime and, in agriculture, less fertile land must be brought into production). MC and AC will be equal at the level of output where AC is minimized. This is a mathematical necessity and intuitive. If you employ the least expensive labor in the initial phase of production, average and marginal cost will decline. Eventually, additional labor will be more costly. For example, if the labor market is at or near full employment, in order to attract additional workers, you must pay higher wages than they are currently earning elsewhere. Thus, the additional (marginal) labor is more costly, and the higher cost increases the overall average as soon as MC exceeds AC. Exhibit 4-7 illustrates the relationship between AC and MC.

Now combine the revenue and cost functions from Exhibits 4-6 and 4-7. In short-run equilibrium, the perfectly competitive firm can earn an economic profit (or an economic loss). In this example, the equilibrium price, 5.71, is higher than the minimum AC. The firm will always maximize profit at an output level where MR = MC. Recall that in perfect competition, the horizontal demand curve is the marginal revenue and average revenue schedules. By setting output at point A in Exhibit 4-8, where MR = MC, the firm will maximize profits. Total revenue is equal to $P \times Q$—in this case, 5.71 times Q_C. Total cost is equal to Q_C times the average cost of producing Q_C, at point B in Exhibit 4-8. The difference between the two areas is economic profit.

3.4. Factors Affecting Long-Run Equilibrium in Perfectly Competitive Markets

In the long run, economic profit will attract other entrepreneurs to the market, resulting in the production of more output. The aggregate supply will increase, shifting the industry supply

[4]Readers who are familiar with calculus will recognize that MC is simply the derivative of total cost with respect to quantity produced.

EXHIBIT 4-7 Individual Firm's Short-Run Cost Schedules

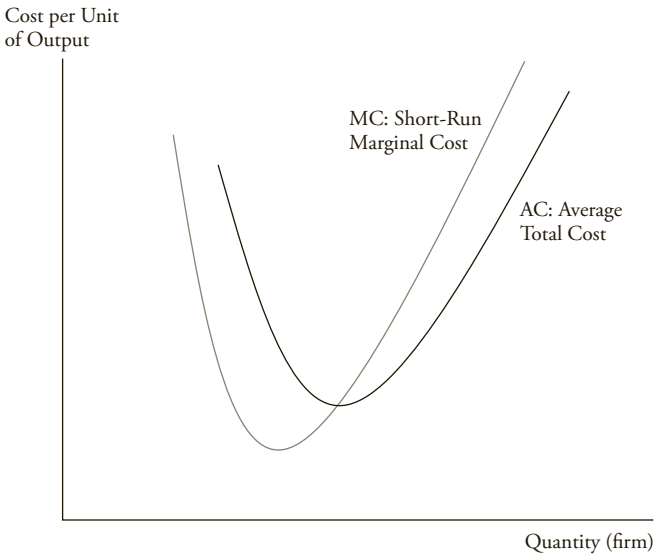


EXHIBIT 4-8 Perfectly Competitive Firm's Short-Run Equilibrium

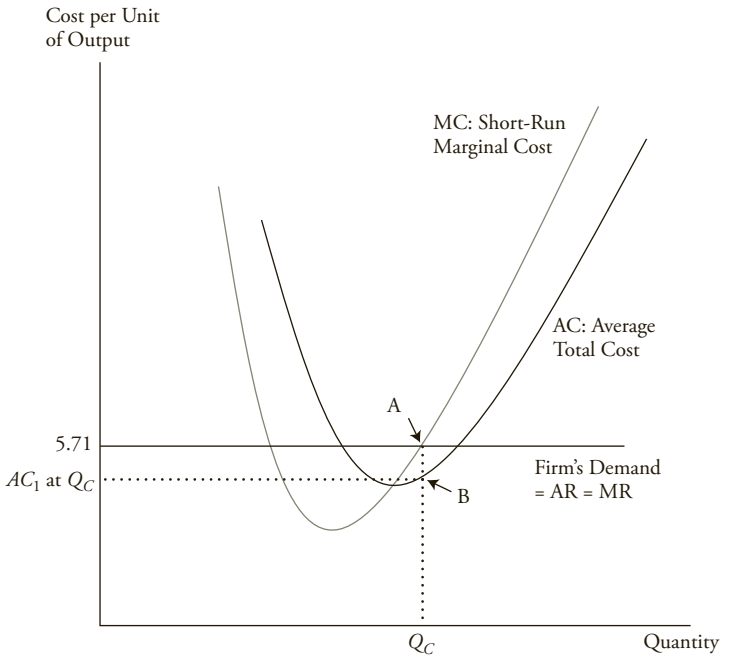
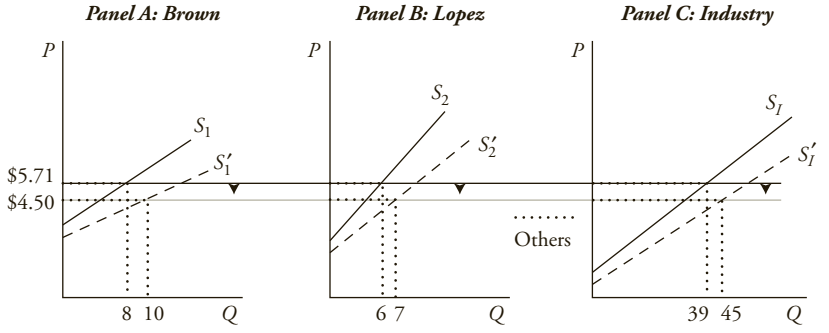


EXHIBIT 4-9 Perfectly Competitive Market with Increased Supply



(S_1) curve to the right, away from the origin of the graph. For a given demand curve, this increase in supply at each price level will lower the equilibrium price, as shown in Exhibit 4-9.

In the long run, the perfectly competitive firm will operate at the point where marginal cost equals the minimum of average cost, because at that point, entry is no longer profitable: In equilibrium, price equals not only marginal cost (firm equilibrium) but also minimum average cost, so that total revenues equal total costs. This result implies that the perfectly competitive firm operates with zero economic profit. That is, the firm receives its normal profit (rental cost of capital), which is included in its economic costs. Recall that economic profits occur when total revenue exceeds total cost (and therefore economic profits differ from accounting profits). With low entry cost and homogeneous products to sell, the perfectly competitive firm earns zero economic profit in the long run.

Exhibit 4-10 illustrates the long-run equilibrium position of the perfectly competitive firm. Note that total revenue equals price ($4.50) times quantity ($Q_E$) and total cost equals average cost ($4.50) times quantity ($Q_E$).

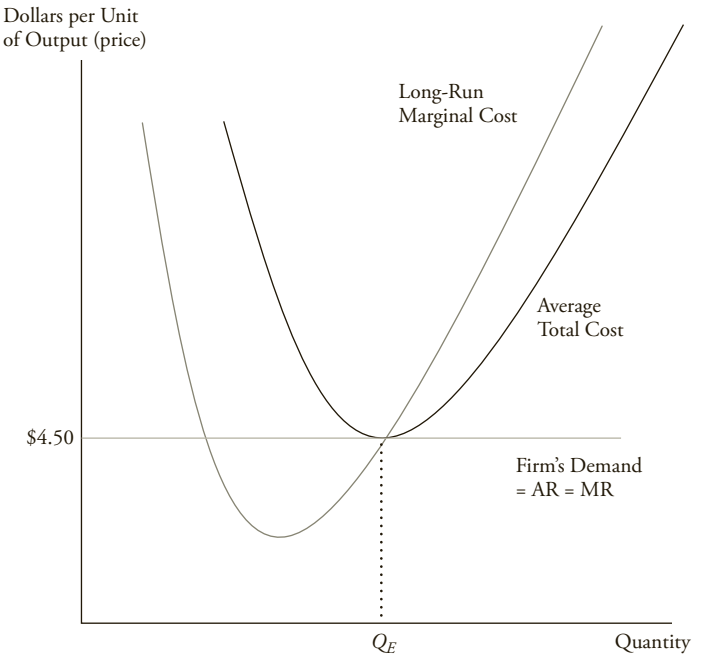
The long-run marginal cost schedule is the perfectly competitive firm's supply curve. The firm's demand curve is dictated by the aggregate market's equilibrium price. The basic rule of profit maximization is that MR = MC, as is the case in long-run equilibrium. The firm's demand schedule is the same as the firm's marginal revenue and average revenue. Given its cost of operation, the only decision the perfectly competitive firm faces is how much to produce. The answer is the level of output that maximizes its return, and that level is where MR = MC. The demand curve is perfectly elastic. Of course, the firm constantly tries to find ways to lower its cost in the long run.

4. MONOPOLISTIC COMPETITION

Early in the twentieth century, economists began to realize that most markets did not operate under the conditions of perfect competition.[5] Many market structures exhibited characteristics of strong competitive forces; however, other distinct noncompetitive factors played important roles in the market. As the name implies, monopolistic competition is a hybrid market. *The*

[5]Chamberlin (1933).

EXHIBIT 4-10 Perfectly Competitive Firm's Long-Run Equilibrium



Schumpeter on Innovation and Perfect Competition

The Austrian-American economist Joseph A. Schumpeter* pointed out that technical change in economics can happen in two main ways:

1. Innovation of process: a new, more efficient way to produce an existing good or service.
2. Innovation of product: a new product altogether or an innovation on an existing product.

Innovation of process is related to production methods. For example, instead of mixing cement by hand, since the invention of the electric engine it has been possible to use electric mixers. A more recent innovation has been to use the Internet to provide technical support to personal computer users: a technician can remotely log on to the customer's PC and fix problems instead of providing instructions over the phone. The result is likely the same, but the process is more efficient.

Innovation of product is related to the product itself. MP3 players, smart phones, robot surgery, and GPS vehicle monitoring have existed for only a few years. They are new products and services. While portable music players existed before the MP3 player,

no similar service existed before GPS monitoring of personal vehicles and freight trucks was invented.

How does the reality of continuous innovation of product and process, which is a characteristic of modern economies, fit into the ideal model of perfect competition, where the product is made by a huge number of tiny, anonymous suppliers? This seems a contradiction because the tiny suppliers cannot all be able to invent new products—and indeed, the markets for portable music players and smart phones do not look like perfect competition.

Schumpeter suggested that perfect competition is more of a long-run type of market. In the short run, a company develops a new process or product and is the only one to take advantage of the innovation. This company likely will have high profits and will outpace any competitors. A second stage is what Schumpeter called the swarming (as when a group of bees leaves a hive to follow a queen): In this case, some entrepreneurs notice the innovation and follow the innovator through imitation. Some of them will fail, while others will succeed and possibly be more successful than the initial innovator. The third stage occurs when the new technology is no longer new because everyone has imitated it. At this point, no economic profits are realized, because the new process or product is no longer a competitive advantage; everyone has it—which is when perfect competition prevails and we have long-run equilibrium until a new innovation of process or product is introduced.

*See part 2 of Schumpeter (1942) for the famous "creative destruction" process.

most distinctive factor in monopolistic competition is product differentiation. Recall the five characteristics from Exhibit 4-1:

1. There is a large number of potential buyers and sellers.
2. The products offered by each seller are close substitutes for the products offered by other firms, and each firm tries to make its product look different.
3. Entry into and exit from the market are possible with fairly low costs.
4. Firms have some pricing power.
5. Suppliers differentiate their products through advertising and other nonprice strategies.

While the market is made up of many firms that comprise the product group, each producer attempts to distinguish its product from those of the others. Product differentiation is accomplished in a variety of ways. For example, consider the wide variety of communication devices available today. Decades ago, when each communication market was controlled by a regulated single seller (the telephone company), all telephones were alike. In the deregulated market of today, the variety of physical styles and colors is extensive. All versions accomplish many of the same tasks.

The communication device manufacturers and providers differentiate their products with different colors, styles, networks, bundled applications, conditional contracts, functionality, and more. Advertising is usually the avenue pursued to convince consumers there is a difference between the goods in the product group. Successful advertising and trademark branding result in customer loyalty. A good example is the brand loyalty associated with Harley-Davidson motorcycles. Harley-Davidson's customers believe that their motorcycles are

truly different from and better than all other motorcycles. The same kind of brand loyalty exists for many fashion creations and cosmetics.

The extent to which the producer is successful in product differentiation determines pricing power in the market. Very successful differentiation results in a market structure that resembles the single-seller market (monopoly). However, because there are relatively low entry and exit costs, competition will, in the long run, drive prices and revenues down toward an equilibrium similar to perfect competition. Thus, the hybrid market displays characteristics found in both perfectly competitive and monopoly markets.

4.1. Demand Analysis in Monopolistically Competitive Markets

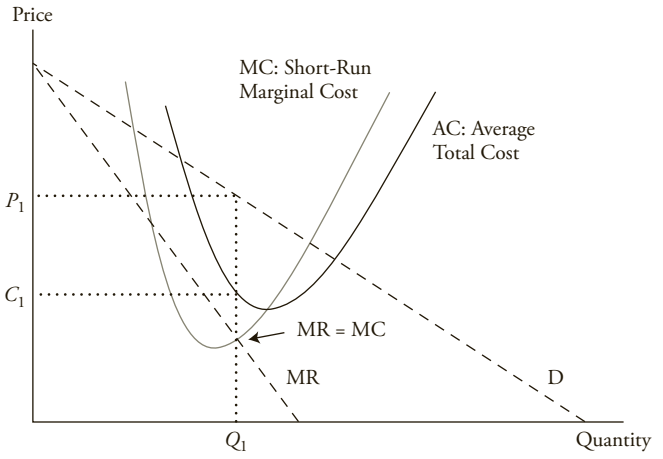
Because each good sold in the product group is somewhat different from the others, the demand curve for each firm in the monopolistic competition market structure is downward sloping to the right. Price and the quantity demanded are negatively related. Lowering the price will increase the quantity demanded, and raising the price will decrease the quantity demanded. There will be ranges of prices within which demand is elastic and (lower) prices at which demand is inelastic. Exhibit 4-11 illustrates the demand, marginal revenue, and cost structures facing a monopolistically competitive firm in the short run.

In the short run, the profit-maximizing choice is the level of output where MR = MC. Because the product is somewhat different from those of the competitors, the firm can charge the price determined by the demand curve. Therefore, in Exhibit 4-11, Q_1 is the ideal level of output and P_1 is the price consumers are willing to pay to acquire that quantity. Total revenue is the area of the rectangle $P_1 \times Q_1$.

4.2. Supply Analysis in Monopolistically Competitive Markets

In perfect competition, the firm's supply schedule is represented by the marginal cost schedule. In monopolistic competition, there is no well-defined supply function. The information used to determine the appropriate level of output is based on the intersection of MC

EXHIBIT 4-11 Short-Run Equilibrium in Monopolistic Competition



and MR. However, the price that will be charged is based on the market demand schedule. The firm's supply curve should measure the quantity the firm is willing to supply at various prices. That information is not represented by either marginal cost or average cost.

4.3. Optimal Price and Output in Monopolistically Competitive Markets

As seen in Section 4.1, in the short run, the profit-maximizing choice is the level of output where MR = MC and total revenue is the area of the rectangle $P_1 \times Q_1$ in Exhibit 4-11.

The average cost of producing Q_1 units of the product is C_1, and the total cost is the area of the rectangle $C_1 \times Q_1$. The difference between TR and TC is economic profit. The profit relationship is described as:

$$\pi = \text{TR} - \text{TC}$$

where π is total profit, TR is total revenue, and TC is total cost.

The Benefits of Imperfect Competition

Is monopolistic competition indeed imperfect—that is, is it a bad thing? At first, one would say that it is an inefficient market structure because prices are higher and the quantity supplied is less than in perfect competition. At the same time, in the real world, we see more markets characterized by monopolistic competition than markets meeting the strict conditions of perfect competition. If monopolistic competition were that inefficient, one wonders, why would it be so common?

A part of the explanation goes back to Schumpeter. Firms try to differentiate their products to meet the needs of customers. Differentiation provides a profit incentive to innovate, experiment with new products and services, and potentially improve the standard of living.

Moreover, because each customer has tastes and preferences that are a bit different, slight variations of each good or service are likely to capture the niche of the market that prefers those variations. An example is the market for candy, where one can find chocolate, licorice, mint, fruit, and many other flavors.

A further reason why monopolistic competition may be good is that people like variety. Traditional economic theories of international trade suggested that countries should buy products from other countries that they cannot produce domestically. Therefore, Norway should buy bananas from a tropical country and sell crude oil in exchange. But this is not the only kind of exchange that happens in reality: For example, Germany imports Honda, Subaru, and Toyota cars from Japan and sells Volkswagen, Porsche, Mercedes, and BMW cars to Japan. In theory, this should not occur because each of the countries produces good cars domestically and does not need to import them. The truth, however (see, for example, Krugman 1989), is that consumers in both countries enjoy variety. Some Japanese drivers prefer to be at the steering wheel of a BMW, whereas others like Hondas; and the same happens in Germany. Variety and product differentiation, therefore, are not necessarily bad things.

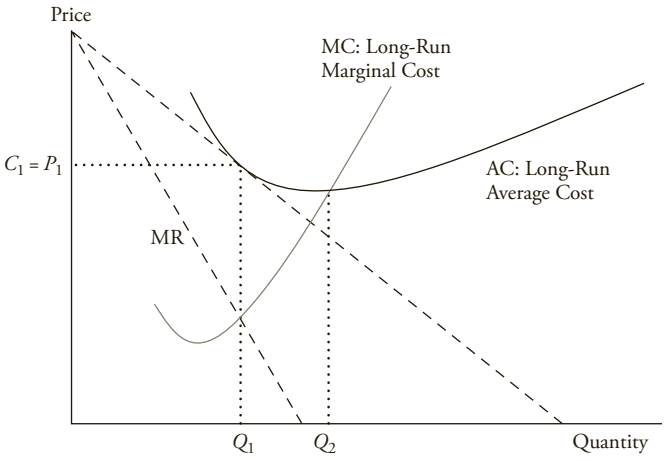
4.4. Factors Affecting Long-Run Equilibrium in Monopolistically Competitive Markets

Because total cost (TC) represents all costs associated with production, including opportunity cost, economic profit is a signal to the market, and that signal will attract more competition. Just as with the perfectly competitive market structure, with relatively low entry costs, more firms will enter the market and lure some customers away from the firm making an economic profit. The loss of customers to new entrant firms will drive down the demand for all firms producing similar products. In the long run for the monopolistically competitive firm, economic profit will fall to zero. Exhibit 4-12 illustrates the condition of long-run equilibrium for monopolistic competition.

In long-run equilibrium, output is still optimal at the level where MR = MC, which is Q_1 in Exhibit 4-12. Again, the price consumers are willing to pay for any amount of the product is determined from the demand curve. That price is P_1 for the quantity Q_1 in Exhibit 4-12, and total revenue is the area of the rectangle $P_1 \times Q_1$. Notice that unlike long-run equilibrium in perfect competition, in the market of monopolistic competition, the equilibrium position is at a higher level of average cost than the level of output that minimizes average cost. Average cost does not reach its minimum until output level Q_2 is achieved. Total cost in this long-run equilibrium position is the area of the rectangle $C_1 \times Q_1$. Economic profit is total revenue minus total cost. In Exhibit 4-12, economic profit is zero because total revenue equals total cost: $P_1 \times Q_1 = C_1 \times Q_1$.

In the hybrid market of monopolistic competition, zero economic profit in long-run equilibrium resembles perfect competition. However, the long-run level of output, Q_1, is less than Q_2, which corresponds to the minimum average cost of production and would be the long-run level of output in a perfectly competitive market. In addition, the economic cost in monopolistic competition includes some cost associated with product differentiation, such as advertising. In perfect competition, there are no costs associated with advertising or marketing because all products are homogeneous. Prices are lower, but consumers may have little variety.

EXHIBIT 4-12 Long-Run Equilibrium in Monopolistic Competition



5. OLIGOPOLY

An oligopoly market structure is characterized by only a few firms doing business in a relevant market. The products must all be similar and, to a great extent, be substitutes for one another. In some oligopoly markets, the goods or services may be differentiated by marketing and strong brand recognition, as in the markets for breakfast cereals and for bottled or canned beverages. Other examples of oligopoly markets are made up of homogeneous products with little or no attempt at product differentiation, such as petroleum and cement. *The most distinctive characteristic of oligopoly markets is the small number of firms that dominate the market. There are so few firms in the relevant market that their pricing decisions are interdependent*. That is, each firm's pricing decision is based on the expected retaliation by the other firms. Recall from Exhibit 4-1 the five characteristics of oligopoly markets:

1. There is a small number of potential sellers.
2. The products offered by each seller are close substitutes for the products offered by other firms and may be differentiated by brand or be homogeneous and unbranded.
3. Entry into the market is difficult, with fairly high costs and significant barriers to competition.
4. Firms typically have substantial pricing power.
5. Products are often highly differentiated through marketing, features, and other nonprice strategies.

Because there are so few firms, each firm can have some degree of pricing power, which can result in substantial profits. Another by-product of the oligopoly market structure is the attractiveness of price collusion. Even without price collusion, a dominant firm may easily become the price maker in the market. Oligopoly markets without collusion typically have the most sophisticated pricing strategies. Examples of noncolluding oligopolies include the U.S. tobacco market and the Thai beer market. In 2004, four firms controlled 99 percent of the U.S. tobacco industry.[6] Brands owned by Singha Co. and by ThaiBev controlled over 90 percent of the Thai beer market in 2009. (This situation is expected to change soon, as the Association of Southeast Asian Nations trade agreement will open the doors to competition from other ASEAN producers.) Perhaps the best-known oligopoly market with collusion is the Organization of Petroleum Exporting Countries (OPEC) cartel, which seeks to control prices in the petroleum market by fostering agreements among oil-producing countries.

5.1. Demand Analysis and Pricing Strategies in Oligopoly Markets

Oligopoly markets' demand curves depend on the degree of pricing interdependence. In a market where collusion is present, the aggregate market demand curve is divided up by the individual production participants. Under noncolluding market conditions, each firm faces an individual demand curve. Furthermore, noncolluding oligopoly market demand characteristics depend on the pricing strategies adopted by the participating firms. There are three basic pricing strategies: pricing interdependence, the Cournot assumption, and the Nash equilibrium.

[6]These examples are based on "Industry Surveys," Net Advantage Database, Standard & Poor's; and Market Share Reports, Gale Research, annual issues, as noted in McGuigan et al. (2008).

The first pricing strategy is to assume pricing interdependence among the firms in the oligopoly. A good example of this situation is any market where there are price wars, such as the commercial airline industry. For example, flying out of the Atlanta, Georgia, hub, Delta Air Lines and AirTran Airways jointly serve several cities. AirTran is a low-cost carrier and typically offers lower fares to destinations out of Atlanta. Delta tends to match the lower fares for those cities also served by AirTran when the departure and arrival times are similar to its own. However, when Delta offers service to the same cities at different time slots, Delta's ticket prices are higher.

The most common pricing strategy assumption in these price war markets is that competitors will match a price reduction and ignore a price increase. The logic is that by lowering its price to match a competitor's price reduction, the firm will not experience a reduction in customer demand. Conversely, by not matching the price increase, the firm stands to attract customers away from the firm that raised its prices. The oligopolist's demand relationship must represent the potential increase in market share when rivals' price increases are not matched and no significant change in market share when rivals' price decreases are matched.

Given a prevailing price, the price elasticity of demand will be much greater if the price is increased and less if the price is decreased. The firm's customers are more responsive to price increases because its rivals have lower prices. Alternatively, the firm's customers are less responsive to price decreases because its rivals will match its price change.

This implies that the oligopolistic firm faces two different demand structures, one associated with price increases and another relating to price reductions. Each demand function will have its own marginal revenue structure as well. Consider the demand and marginal revenue functions in Exhibit 4-13a. The functions $D_{P\uparrow}$ and $MR_{P\uparrow}$ represent the demand and marginal revenue schedules associated with higher prices, while the functions $D_{P\downarrow}$ and $MR_{P\downarrow}$ represent the lower prices' demand and marginal revenue schedules. The two demand schedules intersect at the prevailing price (i.e., the price where price increase and price decrease are both equal to zero).

EXHIBIT 4-13a Kinked Demand Curve in Oligopoly Market: Demand and Marginal Revenue Functions

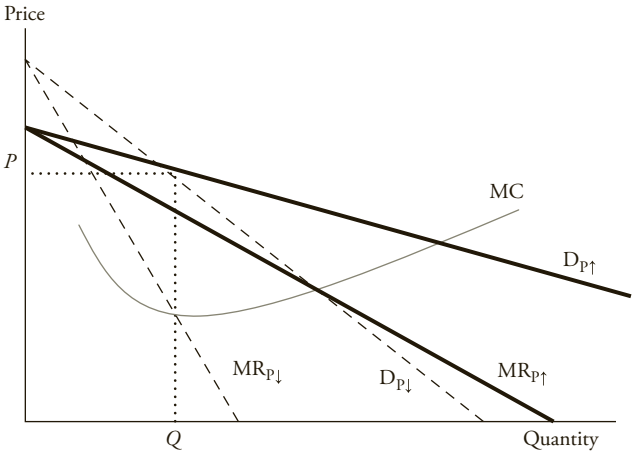
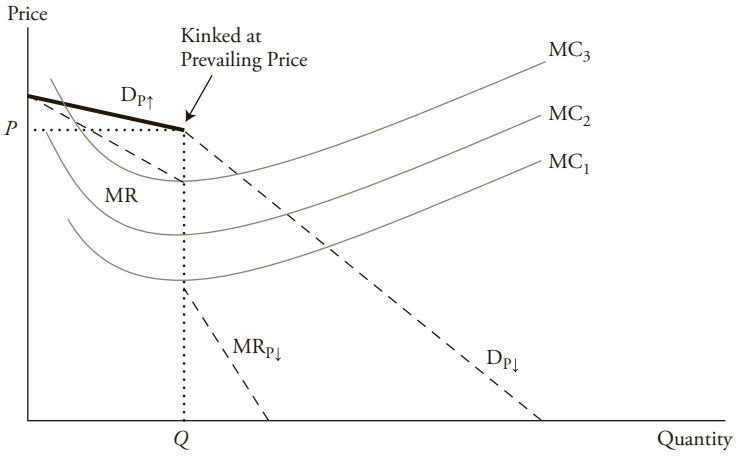


EXHIBIT 4-13b Kinked Demand Curve in Oligopoly Market: Cost Structures



This oligopolistic pricing strategy results in a kinked demand curve, with the two segments representing the different competitor reactions to price changes. The kink in the demand curve also yields a discontinuous marginal revenue structure, with one part associated with the price increase segment of demand and the other relating to the price decrease segment. Therefore, the firm's overall demand equals the relevant portion of $D_{P\uparrow}$ and the relevant portion of $D_{P\downarrow}$. Exhibit 4-13b represents the firm's new demand and marginal revenue schedules. The firm's demand schedule in Exhibit 4-13b is segment $D_{P\uparrow}$ and $D_{P\downarrow}$, where overall demand $D = D_{P\uparrow} + D_{P\downarrow}$.

Notice in Exhibit 4-13b that a wide variety of cost structures are consistent with the prevailing price. If the firm has relatively low marginal costs, MC_1, the profit-maximizing pricing rule established earlier, MR = MC, still holds for the oligopoly firm. Marginal cost can rise to MC_2 and MC_3 before the firm's profitability is challenged. If the marginal cost curve MC_2 passes through the gap in marginal revenue, the most profitable price and output combination remains unchanged at the prevailing price and original level of output.

Criticism of the kinked demand curve analysis focuses on its inability to determine what the prevailing price is from the outset. The kinked demand curve analysis does help explain why stable prices have been observed in oligopoly markets and is therefore a useful tool for analyzing such markets. However, because it cannot determine the original prevailing price, it is considered an incomplete pricing analysis.

The second pricing strategy was first developed by French economist Augustin Cournot in 1838. In the **Cournot assumption**, each firm determines its profit-maximizing production level by assuming that the other firms' outputs will not change. This assumption simplifies pricing strategy because there is no need to guess what the other firms will do to retaliate. It also provides a useful approach to analyzing real-world behavior in oligopoly markets. Take the most basic oligopoly market situation, a two-firm duopoly market.[7] In equilibrium, neither firm has an incentive to change output, given the other firm's production level. Each

[7]The smallest possible oligopoly market is a duopoly, which is made up of only two sellers.

firm attempts to maximize its own profits under the assumption that the other firm will continue producing the same level of output in the future. The Cournot strategy assumes that this pattern continues until each firm reaches its long-run equilibrium position. In long-run equilibrium, output and price are stable: there is no change in price or output that will increase profits for either firm.

Consider this example of a duopoly market. Assume that the aggregate market demand has been estimated to be:

$$Q_D = 450 - P$$

The supply function is represented by constant marginal cost MC = 30.

The Cournot strategy's solution can be found by setting $Q_D = q_1 + q_2$, where q_1 and q_2 represent the output levels of the two firms. Each firm seeks to maximize profit, and each firm believes the other firm will not change output as it changes its own output (Cournot's assumption). The firm will maximize profit where MR = MC. Rearranging the aggregate demand function in terms of price, we get:

$$P = 450 - Q_D = 450 - q_1 - q_2, \text{ and } \mathrm{MC} = 30$$

Total revenue for each of the two firms is found by multiplying price and quantity:

$$\mathrm{TR}_1 = Pq_1 = (450 - q_1 - q_2)q_1 = 450q_1 - q_1^2 - q_1q_2, \text{ and}$$
$$\mathrm{TR}_2 = Pq_2 = (450 - q_1 - q_2)q_2 = 450q_2 - q_2q_1 - q_2^2$$

Marginal revenue is defined as the change in total revenue, given a change in sales (q_1 or q_2).[8] For the profit-maximizing output, set MR = MC, or:

$$450 - 2q_1 - q_2 = 30$$

and:

$$450 - q_1 - 2q_2 = 30$$

Find the simultaneous equilibrium for the two firms by solving the two equations with two unknowns:

$$450 - 2q_1 - q_2 = 450 - q_1 - 2q_2$$

Because $q_2 = q_1$ under Cournot's assumption, insert this solution into the demand function and solve as:

$$450 - 2q_1 - q_1 = 450 - 3q_1 = 30$$

Therefore, $q_1 = 140$, $q_2 = 140$, and $Q = 280$. The price is $P = 450 - 280 = 170$.

[8]The marginal revenue formulas can be obtained using the technique introduced in Section 3.1. For the market demand function, total revenue is $P \times Q = 450Q - Q_2$, and our technique yields $\mathrm{MR} = \Delta\mathrm{TR}/\Delta Q = 450 - 2Q$. For the individual firms in the Cournot duopoly, $\mathrm{MR}_1 = \Delta\mathrm{TR}_1/\Delta q_1 = 450 - 2q_1 - q_2$, and $\mathrm{MR}_2 = \Delta\mathrm{TR}_2/\Delta q_2 = 450 - q_1 - 2q_2$. Each of these marginal revenue formulas is, of course, the derivative of the relevant total revenue formula with respect to the relevant quantity.

In the Cournot strategic pricing solution, the market equilibrium price will be 170 and the aggregate output will be 280 units. This result, known as the Cournot equilibrium, differs from the perfectly competitive market equilibrium because the perfectly competitive price will be lower and the perfectly competitive output will be higher. In general, non-competitive markets have higher prices and lower levels of output in equilibrium than do markets with perfect competition. In competition, the equilibrium is reached where price equals marginal cost.

$$P_C = \text{MR}_C = \text{MC}, \text{ so } 450 - Q = 30$$

where P_C is the competitive firm's equilibrium price.

$$Q = 420, \text{ and } P_C = 30$$

Exhibit 4-14 describes the oligopoly, competitive, and monopoly market equilibrium positions, where P_M is the monopoly optimum price, P_C is the competitive price, and $P_{Cournot}$ is the oligopoly price under the Cournot assumption.

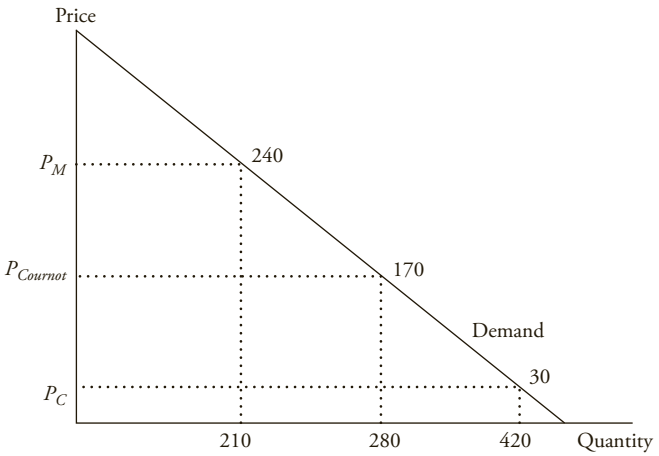
In the later discussion regarding monopoly market structure, equilibrium will be established where MR = MC. That solution is also shown in Exhibit 4-14. The monopoly firm's demand schedule is the aggregate market demand schedule. Therefore, the solution is:

$$\text{MR} = \text{MC}$$

From footnote 8, MR = 450 − 2*Q*; therefore,

$$450 - 2Q = 30, \text{ and } Q = 210$$

EXHIBIT 4-14 Cournot Equilibrium in Duopoly Market



From the aggregate demand function, solve for price:

$$P_M = 450 - 210 = 240$$

Note that the Cournot solution falls between the competitive equilibrium and the monopoly solution.

It can be shown that as the number of firms increases from two to three, from three to four, and so on, the output and price equilibrium positions move toward the competitive equilibrium solution. This result has historically been the theoretical basis for the antitrust policies established in the United States.

The third pricing strategy is attributed to one of the 1994 Nobel Prize winners, John Nash, who first developed the general concepts. In the previous analysis, the concept of market equilibrium occurs when firms are achieving their optimum remuneration under the circumstances they face. In this optimum environment, the firm has no motive to change price or output level. Existing firms are earning a normal return (zero economic profit), leaving no motive for entry to or exit from the market. All firms in the market are producing at the output level where price equals the average cost of production.

In **game theory** (the set of tools that decision makers use to consider responses by rival decision makers), the **Nash equilibrium** is present when two or more participants in a noncooperative game have no incentive to deviate from their respective equilibrium strategies after they have considered and anticipated their opponent's rational choices or strategies. In the context of oligopoly markets, the Nash equilibrium is an equilibrium defined by the characteristic that none of the oligopolists can increase its profits by unilaterally changing its pricing strategy. The assumption is made that each participating firm does the best it can, given the reactions of its rivals. Each firm anticipates that the other firms will react to any change made by competitors by doing the best they can under the altered circumstances. The firms in the oligopoly market have interdependent actions. The actions are noncooperative, with each firm making decisions that maximize its own profits. The firms do not collude in an effort to maximize joint profits. The equilibrium is reached when all firms are doing the best they can, given the actions of their rivals.

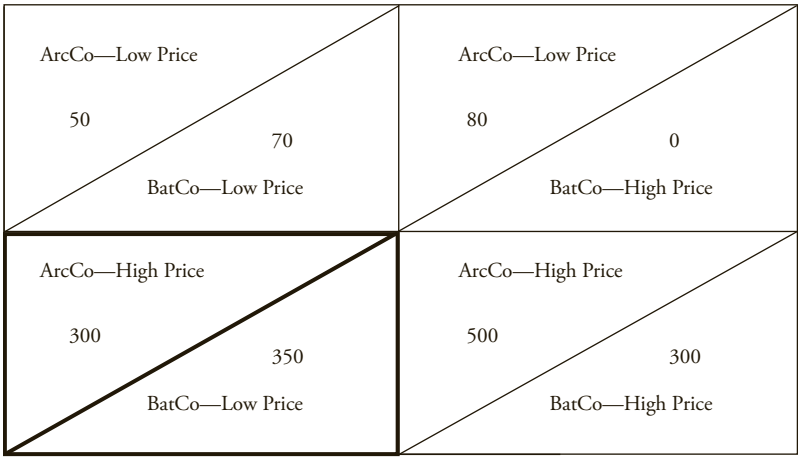
Exhibit 4-15 illustrates the duopoly result from the Nash equilibrium. Assume there are two firms in the market, ArcCo and BatCo. ArcCo and BatCo can charge high prices or low prices for the product. The market outcomes are shown in Exhibit 4-15.

For example, the top left solution indicates that when both ArcCo and BatCo offer the product at low prices, ArcCo earns a profit of 50 and BatCo earns 70. The top right solution shows that if ArcCo offers the product at a low price and BatCo offers the product at a high price, BatCo earns zero profits. The solution with the maximum joint profits is the lower right equilibrium, where both firms charge high prices for the product. Joint profits are 800 in this solution.

However, the Nash equilibrium requires that each firm behaves in its own best interest. BatCo can improve its position by offering the product at a low price when ArcCo is charging a high price. In the lower left solution, BatCo maximizes its profits at 350. While ArcCo can earn 500 in its best solution, it can do so only if BatCo also agrees to charge a high price. This option is clearly not in BatCo's best interest because it can increase its return from 300 to 350 by charging a low price if ArcCo is charging a high price.

This scenario brings up the possibility of collusion. If ArcCo agrees to share at least 51 of its 500 when both companies are charging high prices, BatCo should also be willing to charge high prices. While, in general, such collusion is unlawful in most countries, it remains a

EXHIBIT 4-15 Nash Equilibrium in Duopoly Market



tempting alternative. Clearly, conditions in oligopolistic industries encourage collusion, with a small number of competitors and interdependent pricing behavior. Collusion is motivated by several factors: increased profits, reduced cash flow uncertainty, and improved opportunities to construct barriers to entry.

When collusive agreements are made openly and formally, the firms involved are called a **cartel**. In some cases, collusion is successful; other times, the forces of competition overpower collusive behavior. There are six major factors that affect the chances of successful collusion.[9]

1. *The number and size distribution of sellers.* Successful collusion is more likely if the number of firms is small or if one firm is dominant. Collusion becomes more difficult as the number of firms increases or if the few firms have similar market shares. When the firms have similar market shares, the competitive forces tend to overshadow the benefits of collusion.
2. *The similarity of the products.* When the products are homogeneous, collusion is more successful. The more differentiated the products, the less likely it is that collusion will succeed.
3. *Cost structure.* The more similar the firms' cost structures, the more likely it is that collusion will succeed.
4. *Order size and frequency.* Successful collusion is more likely when orders are frequent, received on a regular basis, and relatively small. Frequent small orders, received regularly, diminish the opportunities and rewards for cheating on the collusive agreement.
5. *The strength and severity of retaliation.* Oligopolists will be less likely to break the collusive agreement if the threat of retaliation by the other firms in the market is severe.
6. *The degree of external competition.* The main reason to enter into the formal collusion is to increase overall profitability of the market, and rising profits attract competition. For example, the average extraction cost of a barrel of crude oil from the Persian Gulf is $3, while the average cost from the Alaskan fields is $30. It is more likely that crude oil

[9]McGuigan et al. (2008).

producers in the Persian Gulf will successfully collude because of the similarity in their cost structures. If OPEC had held crude oil prices down below $30 per barrel, there would not have been a viable economic argument to explore oil fields in Alaska. Extracting petroleum from Canadian tar sands becomes economically attractive only at prices above $50 per barrel. OPEC's successful cartel raised crude oil prices to the point where outside sources became economically possible and in doing so increased the competition the cartel faces.

There are other possible oligopoly strategies that are associated with decision making based on game theory. The Cournot equilibrium and the Nash equilibrium are examples of specific strategic games. A strategic game is any interdependent behavioral choice employed by individuals or groups that share a common goal (e.g., military units, sports teams, or business decision makers). Another prominent decision-making strategy in oligopolistic markets is the first-mover advantage in the **Stackelberg model**, named after the economist who first conceptualized the strategy.[10] The important difference between the Cournot model and the Stackelberg model is that Cournot assumes that in a duopoly market, decision making is simultaneous, while Stackelberg assumes that decisions are made sequentially. In the Stackelberg model, the leader firm chooses its output first and then the follower firm chooses after observing the leader's output. It can be shown that the leader firm has a distinct advantage, being a first mover.[11] In the Stackelberg game, the leader can aggressively overproduce to force the follower to scale back its production or even punish or eliminate the weaker opponent. This approach is sometimes referred to as a "top dog" strategy.[12] The leader earns more than in Cournot's simultaneous game, while the follower earns less. Many other strategic games are possible in oligopoly markets. The important conclusion is that the optimal strategy of the firm depends on what its adversary does. The price and marginal revenue the firm receives for its product depend on both its decisions and its adversary's decisions.

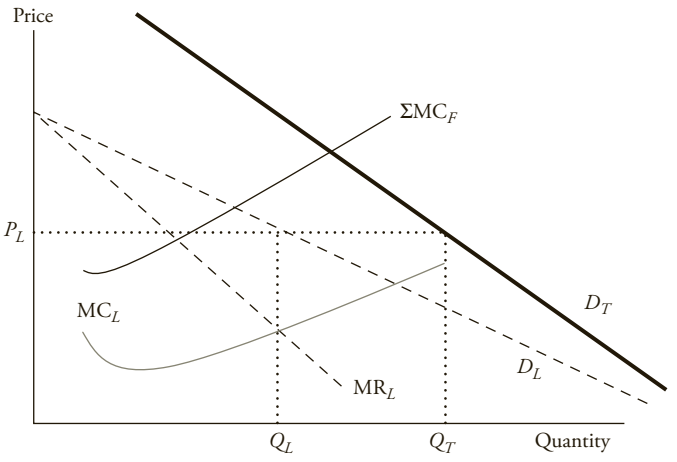
5.2. Supply Analysis in Oligopoly Markets

As in monopolistic competition, the oligopolist does not have a well-defined supply function. That is, there is no way to determine the oligopolist's optimal levels of output and price independent of demand conditions and competitor's strategies. However, the oligopolist still has a cost function that determines the optimal level of supply. Therefore, the profit-maximizing rule established earlier is still valid: the level of output that maximizes profit is where MR = MC. The price to charge is determined by what price consumers are willing to pay for that quantity of the product. Therefore, the equilibrium price comes from the demand curve, while the output level comes from the relationship between marginal revenue and marginal cost.

Consider an oligopoly market in which one of the firms is dominant and thus able to be the price leader. Dominant firms generally have 40 percent or greater market share. When one firm dominates an oligopoly market, it does so because it has greater capacity, has a lower cost structure, was first to market, or has greater customer loyalty than other firms in the market.

[10]Von Stackelberg (1952). See also Kelly (2003, 115–120), for a comparison between the Cournot and Stackelberg equilibriums.
[11]Nicholson and Snyder (2008, 543).
[12]Fudenberg and Tirole (1984, 361–368).

EXHIBIT 4-16 Dominant Oligopolist's Price Leadership



Assuming there is no collusion, the dominant firm becomes the price maker, and therefore its actions are similar to monopoly behavior in its segment of the market. The other firms in the market follow the pricing patterns of the dominant firm. Why wouldn't the price followers attempt to gain market share by undercutting the dominant firm's price? The most common explanation is that the dominant firm's supremacy often stems from a lower cost of production. Usually, the price followers would rather charge a price that is even higher than the dominant firm's price choice. If they attempt to undercut the dominant firm, the followers risk a price war with a lower-cost producer that can threaten their survival. Some believe that one explanation for the price leadership position of the dominant firm is simply convenience. Only one firm has to make the pricing decisions, and the others can simply follow its lead.

Exhibit 4-16 establishes the dominant firm's pricing decision. The dominant firm's demand schedule, D_L, is a substantial share of the total market demand, D_T. The low-cost position of the dominant firm is represented by its marginal cost, MC_L. The sum of the marginal costs of the price followers is established as ΣMC_F and represents a higher cost of production than that of the price leader.

There is an important reason why the total demand curve and the leader demand curve are not parallel in Exhibit 4-16: Remember that the leader is the low-cost producer. Therefore, as price decreases, fewer of the smaller suppliers will be able to profitably remain in the market, and several will exit because they do not want to sell below cost. Therefore, the leader will have a larger market share as P decreases, which implies that Q_L increases at a low price, exactly as shown by a steeper D_T in the diagram.

The price leader identifies its profit-maximizing output where $MR_L = MC_L$, at output Q_L. This is the quantity it wants to supply; however, the price it will charge is determined by its segment of the total demand function, D_L. At price P_L, the dominant firm will supply quantity Q_L of total demand, D_T. The price followers will supply the difference to the market, $(Q_T - Q_L) = Q_F$. Therefore, neither the dominant firm nor the follower firms have a single functional relationship that determines the quantity supplied at various prices.

5.3. Optimal Price and Output in Oligopoly Markets

From the preceding discussion, it is clear that there is no single optimum price and output analysis that fits all oligopoly market situations. The interdependence among the few firms that make up the oligopoly market provides a complex set of pricing alternatives, depending on the circumstances in each market. In the case of the kinked demand curve, the optimum price is the prevailing price at the kink in the demand function. However, as previously noted, the kinked demand curve analysis does not provide insight into what established the prevailing price in the first place.

Perhaps the case of the dominant firm, with the other firms following the price leader, is the most obvious. In that case, the optimal price is determined at the output level where MR = MC. The profit-maximizing price is then determined by the output position of the segment of the demand function faced by the dominant firm. The price followers have little incentive to change the leader's price. In the case of the Cournot assumption, each firm assumes that the other firms will not alter their outputs following the dominant firm's selection of its price and output level.

Therefore, again, the optimum price is determined by the output level where MR = MC. In the case of the Nash equilibrium, each firm will react to the circumstances it faces, maximizing its own profit. These adjustments continue until there are stable prices and levels of output. Because of the interdependence, there is no certainty as to the individual firm's price and output level.

5.4. Factors Affecting Long-Run Equilibrium in Oligopoly Markets

Long-run economic profits are possible for firms operating in oligopoly markets. However, history has shown that, over time, the market share of the dominant firm declines. Profits attract entry by other firms into the oligopoly market. Over time, the marginal costs of the entrant firms decrease because they adopt more efficient production techniques, the dominant firm's demand and marginal revenue shrink, and the profitability of the dominant firm declines. In the early 1900s, J. P. Morgan, Elbert Gary, Andrew Carnegie, and Charles M. Schwab created the United States Steel Corporation (U.S. Steel). When it was first formed in 1901, U.S. Steel controlled 66 percent of the market. By 1920, U.S. Steel's market share had declined to 46 percent, and by 1925 its market share was 42 percent.

In the long run, optimal pricing strategy must include the reactions of rival firms. History has proven that pricing wars should be avoided because any gains in market share are temporary. Decreasing prices to drive competitors away lowers total revenue to all participants in the oligopoly market. Innovation may be a way—though sometimes an uneconomical one—to maintain market leadership.

Oligopolies: Appearance versus Behavior

When is an oligopoly not an oligopoly? There are two extreme cases of this situation. A normal oligopoly has a few firms producing a differentiated good, and this differentiation gives them pricing power.

At one end of the spectrum, we have the oligopoly with a credible threat of entry. In practice, if the oligopolists are producing a good or service that can be easily

replicated, has limited economies of scale, and is not protected by brand recognition or patents, they will not be able to charge high prices. The easier it is for a new supplier to enter the market, the lower the margins. In practice, this oligopoly will behave very much like a perfectly competitive market.

At the opposite end of the spectrum, we have the case of the cartel. Here, the oligopolists collude and act as if they were a single firm. In practice, a very effective cartel enacts a cooperative strategy. As shown in Section 5.1, instead of going to a Nash equilibrium, the cartel participants go to the more lucrative (for them) cooperative equilibrium.

A cartel may be explicit (that is, based on a contract) or implicit (based on signals). An example of signals in a duopoly would be that one of the firms reduces its prices and the other does not. Because the firm not cutting prices refuses to start a price war, the firm that cut prices may interpret this signal as a suggestion to raise prices to a higher level than before, so that profits may increase for both.

6. MONOPOLY

Monopoly market structure is at the opposite end of the spectrum from perfect competition. For various reasons, there are significant barriers to entry such that a single firm produces a highly specialized product and faces no threat of competition. There are no good substitutes for the product in the relevant market, and the market demand function is the same as the individual firm's demand schedule. *The distinguishing characteristics of monopoly are that a single firm represents the market and significant barriers to entry exist.* Exhibit 4-1 identified the five characteristics of monopoly markets:

1. There is a single seller of a highly differentiated product.
2. The product offered by the seller has no close substitute.
3. Entry into the market is very difficult, with high costs and significant barriers to competition.
4. The firm has considerable pricing power.
5. The product is differentiated through nonprice strategies such as advertising.

Monopoly markets are unusual. With a single seller dominating the market, power over price decisions is significant. For a single seller to achieve this power, there must be factors that allow the monopoly to exist. One obvious source of monopoly power would be a patent or copyright that prevents other firms from entering the market. Patent and copyright laws exist to reward intellectual capital and investment in research and development. In so doing, they provide significant barriers to entry.

Another possible source of market power is control over critical resources used for production. One example is De Beers Consolidated Mines Limited. De Beers owned or controlled all diamond-mining operations in South Africa and established pricing agreements with other important diamond producers. In doing so, De Beers was able to control the prices for cut diamonds for decades. Technically, De Beers was a near-monopoly dominant firm rather than a pure monopoly, although its pricing procedure for cut diamonds resembled monopoly behavior.

Perhaps the most common form of monopolistic market power occurs as the result of government-controlled authorization. In most urban areas, a single source of water and sewer services is offered. In some cases, these services are offered by a government-controlled entity. In other cases, private companies provide the services under government regulation. Such so-called natural monopolies require a large initial investment that benefits from economies of scale; therefore, government may authorize a single seller to provide a certain service because having multiple sellers would be too costly. For example, electricity in most markets is provided by a single seller. Economies of scale result when significant capital investment benefits from declining long-run average costs. In the case of electricity, a large gas-fueled power plant producing electricity for a large area is substantially more efficient than having a small diesel generator for every building. That is, the average cost of generating and delivering a kilowatt of electricity will be substantially lower with the single power station, but the initial fixed cost of building the power station and the lines delivering electricity to each home, factory, and office will be very high.

In the case of natural monopolies, limiting the market to a single seller is considered beneficial to society. One water and sewer system is deemed better for the community than dozens of competitors because building multiple infrastructures for running water and sewer service would be particularly expensive and complicated. One electrical power grid supplying electricity for a community can make large capital investments in generating plants and lower the long-run average cost, while multiple power grids would lead to a potentially dangerous maze of wires. Clearly, not all monopolies are in a position to make significant economic profits. Regulators, such as public utility commissions in the United States, attempt to determine what a normal return for the monopoly owners' investment should be, and they set prices accordingly. Nevertheless, monopolists attempt to maximize profits.

Not all monopolies originate from natural barriers. For some monopolists, barriers to entry do not derive from increasing returns to scale. We mentioned that marketing and brand loyalty are sources of product differentiation in monopolistic competition. In some highly successful cases, strong brand loyalty can become a formidable barrier to entry. For example, if the Swiss watchmaker Rolex is unusually successful in establishing brand loyalty so that its customers think there is no close substitute for its product, then the company will have monopoly-like pricing power over its market.

The final potential source of market power is the increasing returns associated with network effects. Network effects result from synergies related to increasing market penetration. By achieving a critical level of adoption, Microsoft was able to extend its market power through the network effect—for example, because most computer users know how to use Microsoft Word. Therefore, for firms, Word is cheaper to adopt than other programs because almost every new hire will be proficient in using the software and will need no further training. At some level of market share, a network-based product or service (think of Facebook or eBay) reaches a point where each additional share point increases the probability that another user will adopt.[13] These network effects increase the value to other potential adopters. In Microsoft's case, the network effects crowded out other potential competitors, including Netscape's Internet browser, that might have led to applications bypassing Windows. Eventually, Microsoft's operating system's market share reached 92 percent of the global market. Similar situations occur in financial markets: If a publicly listed share or a derivative contract is more frequently traded on a certain exchange, market participants wishing to sell or buy the

[13]When a network-based device reaches a 30 percent share, the next 50 percentage points are cheaper to promote, according to McGuigan et al. (2008).

security will go to the more liquid exchange because they expect to find a better price and faster execution there.

6.1. Demand Analysis in Monopoly Markets

The monopolist's demand schedule is the aggregate demand for the product in the relevant market. Because of the income effect and the substitution effect, demand is negatively related to price, as usual. The slope of the demand curve is negative and therefore downward sloping. The general form of the demand relationship is:

$$Q_D = a - bP \quad \text{or, rewritten,} \quad P = a/b - (1/b)Q_D$$

Therefore, total revenue = TR = $P \times Q = (a/b)Q_D - (1/b)Q_D^2$.

Marginal revenue is the change in revenue given a change in the quantity demanded. Because an increase in quantity requires a lower price, the marginal revenue schedule is steeper than the demand schedule. If the demand schedule is linear, then the marginal revenue curve is twice as steep as the demand schedule.[14]

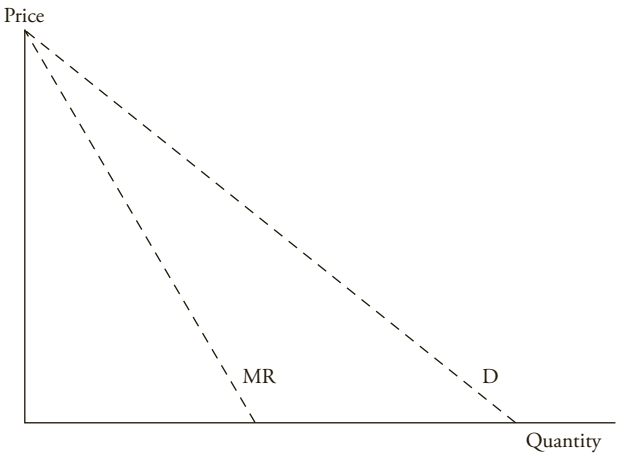
$$\text{MR} = \Delta\text{TR}/\Delta Q = (a/b) - (2/b)Q_D$$

The demand and marginal revenue relationship is expressed in Exhibit 4-17.

Suppose a company operating on a remote island is the single seller of natural gas. Demand for its product can be expressed as:

$$Q_D = 400 - 0.5P, \text{ which can be rearranged as}$$
$$P = 800 - 2Q_D$$

EXHIBIT 4-17 Monopolist's Demand and Marginal Revenue



[14]Marginal revenue can be found using the technique shown in Section 3.1 or, for readers who are familiar with calculus, by taking the derivative of the total revenue function: MR = $\Delta\text{TR}/\Delta Q = (a/b) - (2/b)Q_D$.

Total revenue is $P \times Q = \text{TR} = 800Q_D - 2Q_D^2$, and marginal revenue is $\text{MR} = 800 - 4Q_D$.[15]

In Exhibit 4-17, the demand curve's intercept is 800 and the slope is −2. The marginal revenue curve in Exhibit 4-17 has an intercept of 800 and a slope of −4.

Average revenue is TR/Q_D; therefore, $\text{AR} = 800 - 2Q_D$, which is the same as the demand function. In the monopoly market model, average revenue is the same as the market demand schedule.

6.2. Supply Analysis in Monopoly Markets

A monopolist's supply analysis is based on the firm's cost structure. As in the market structures of monopolistic competition and oligopoly, the monopolist does not have a well-defined supply function that determines the optimal output level and the price to charge. The optimal output is the profit-maximizing output level. The profit-maximizing level of output occurs where marginal revenue equals marginal cost, MR = MC.

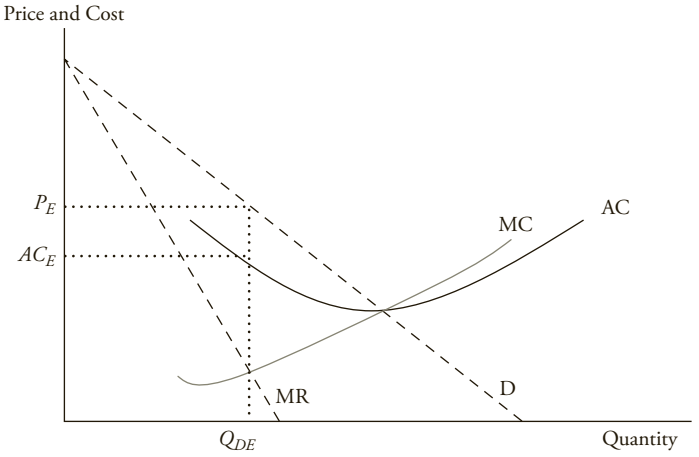
Assume the natural gas company has determined that its total cost can be expressed as:

$$\text{TC} = 20{,}000 + 50Q + 3Q^2$$

Marginal cost is $\Delta\text{TC}/\Delta Q = \text{MC} = 50 + 6Q$.[16]

Supply and demand can be combined to determine the profit-maximizing level of output. Exhibit 4-18 combines the monopolist's demand and cost functions.

EXHIBIT 4-18 Monopolist's Demand, Marginal Revenue, and Cost Structures



[15]$\text{MR} = \Delta\text{TR}/\Delta Q = 800 - 4Q$; see footnote 14.

[16]The marginal cost equation can be found in this case by applying the technique used to find the marginal revenue equation in Section 3.1, or by taking the derivative of the total cost function.

In Exhibit 4-18, the demand and marginal revenue functions are clearly defined by the aggregate market. However, the monopolist does not have a supply curve. The quantity that maximizes profit is determined by the intersection of MC and MR, Q_{DE}.

The price consumers are willing to pay for this level of output is P_E, as determined by the demand curve, P_E.

The profit-maximizing level of output is MR = MC: $800 - 4Q_D = 50 + 6Q_D$; therefore, $Q_D = 75$ when profit is maximized.

Total profit equals total revenue minus total cost:

$$\pi = 800Q - 2Q_D^2 - (20{,}000 + 50Q_D + 3Q_D^2) = -20{,}000 + 750Q_D - 5Q_D^2$$

Profit is represented by the difference between the area of the rectangle $Q_{DE} \times P_E$, representing total revenue, and the area of the rectangle $Q_{DE} \times AC_E$, representing total cost.

Monopolists and Their Incentives

In theoretical models, which usually take product quality and technology as given, monopolists can choose to vary either price or quantity. In real life, they also have the ability to vary their product.

A monopolist can choose to limit quality if producing a higher-quality product is costly and higher quality does not increase profits accordingly. For example, the quality of domestically produced cars in most developed countries improved dramatically once foreign imports became more available. Before the opening of borders to foreign imports, the single incumbent that dominated the market (for example, Fiat in Italy) or the small group of incumbents acting as a collusive oligopoly (such as the Detroit Big Three in the United States) were the effective monopolists of their domestic automobile markets. Rust corrosion, limited reliability, and poor gas mileage were common.*

Similarly, regulated utilities may have limited incentives to innovate. Several studies, including Gomez-Ibanez (2003), have found that state-owned and other monopoly telephone utilities tended to provide very poor service before competition was introduced. Poor service may not be limited to poor connection quality but may also include extensive delays in adding new users and limited introduction of new services, such as caller ID or automatic answering services.

Intuitively, a monopolist will not spend resources on quality control, research and development, or even customer relations unless there is a threat of entry of a new competitor or there is a clear link between such expenses and a profit increase. In contrast, in competitive markets (including oligopoly), innovation and quality are often ways to differentiate the product and increase profits.

*For more on this topic, see Banker, Khosla, and Sinha (1998).

6.3. Optimal Price and Output in Monopoly Markets

Continuing the natural gas example, the total profit function can be solved using the quadratic formula.[17] Another method to solve the profit function is to evaluate $\Delta\pi/\Delta Q_D$ and set it equal to zero. This identifies the point at which profit is unaffected by changes in output.[18] Of course, this will give the same result as we found by equating marginal revenue with marginal cost. The monopoly will maximize profits when $Q^* = 75$ units of output and the price is set from the demand curve at 650.

$$P* = 800 - 2(75) = 650 \text{ per unit}$$

To find total maximum profits, substitute these values into the profit function:

$$\pi = -20{,}000 + 750Q_D - 5Q_D^2 = -20{,}000 + 750(75) - 5(75^2) = 8{,}125$$

Note that the price and output combination that maximizes profit occurs in the elastic portion of the demand curve in Exhibit 4-18. This must be so because marginal revenue and marginal cost will always intersect where marginal revenue is positive. This fact implies that quantity demanded responds more than proportionately to price changes (i.e., demand is elastic) at the point at which MC = MR. As noted earlier, the relationship between marginal revenue and price elasticity, E_P, is:

$$\text{MR} = P(1 - 1/E_P)$$

In monopoly, MR = MC; therefore,

$$P(1 - 1/E_P) = \text{MC}$$

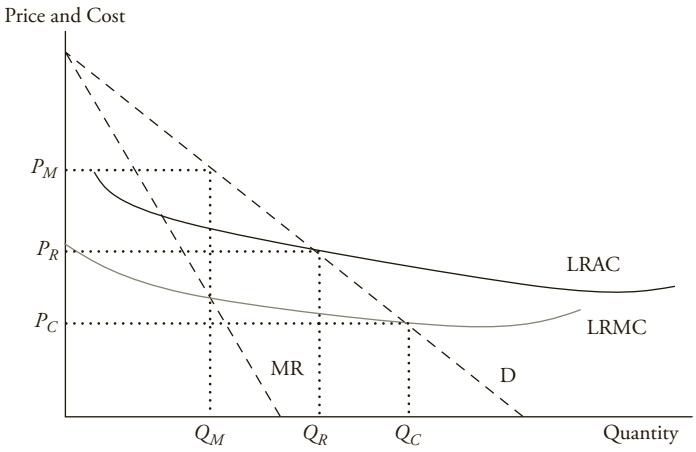
The firm can use this relationship to determine the profit-maximizing price if the firm knows its cost structure and the price elasticity of demand, E_P. For example, assume the firm knows that its marginal cost is constant at 75 and recent market analysis indicates that price elasticity is estimated to be 1.5. The optimal price is solved as:

$$P(1 - 1/1.5) = 75 \quad \text{and} \quad P = 225$$

Exhibit 4-18 indicates that the monopolist wants to produce at Q_E and charge the price of P_E. Suppose this is a natural monopoly that is operating as a government franchise under regulation. Natural monopolies are usually found where production is based on significant economies of scale and declining cost structure in the market. Examples include electric power generation, natural gas distribution, and the water and sewer industries. These are often called public utilities. Exhibit 4-19 illustrates such a market in long-run equilibrium.

[17]The quadratic formula, where $aQ^2 + bQ + c = 0$, is $Q = \{-b \pm \sqrt{(b^2 - 4ac)}\}/2a$.
[18]Maximum profit occurs where $\Delta\pi/\Delta Q_D = 0 = 750 - 10Q_D$. Therefore, profits are maximized at $Q_D = 75$.

EXHIBIT 4-19 Natural Monopoly in a Regulated Pricing Environment



In Exhibit 4-19, three possible pricing and output solutions are presented. The first is what the monopolist would do without regulation: The monopolist would seek to maximize profits by producing Q_M units of the product, where long-run marginal cost equals marginal revenue, LRMC = MR. To maximize profits, the monopolist would raise the price to the level the demand curve will accept, P_M.

In perfect competition, the price and output equilibrium occurs where price is equal to the marginal cost of producing the incremental unit of the product. In a competitive market, the quantity produced would be higher, Q_C, and the price lower, P_C. For this regulated monopoly, the competitive solution would be unfair because at output Q_C, the price P_C would not cover the average cost of production. One possibility is to subsidize the difference between the long-run average cost, LRAC, and the competitive price, P_C, for each unit sold.

Another solution is for the regulator to set the price at the point where long-run average cost equals average revenue. Recall that the demand curve represents the average revenue the firm receives at each output level. The government regulator will attempt to determine the monopolistic firm's long-run average cost and set the output and price so that the firm receives a fair return on the owners' invested capital. The regulatory solution is output level Q_R, with the price set at P_R. Therefore, the regulatory solution is found between the unregulated monopoly equilibrium and the competitive equilibrium.

6.4. Price Discrimination and Consumer Surplus

Monopolists can be more or less effective in taking advantage of their market structure. At one extreme, we have a monopolist that charges prices and supplies quantities that are the same as they would be in perfect competition; this scenario may be a result of regulation or threat of entry (if the monopolist charged more, another company could come in and price the former monopolist out of the market). At the opposite extreme, hated by all consumers and economists, is the monopolist that extracts the entire consumer surplus. This scenario is called **first-degree price discrimination**, where a monopolist is able to charge each customer the

highest price the customer is willing to pay. This is called price discrimination because the monopolist charges a different price to each client. How can this be? For example, if the monopolist knows the exact demand schedule of the customer, then the monopolist is able to capture the entire consumer surplus. In practice, the monopolist is able to measure how often the product is used and charges the customer the highest price the consumer is willing to pay for that unit of good. Another possibility is that public price disclosure is nonexistent, so that no customer knows what the other customers are paying. Interestingly, not every consumer is worse off in this case, because some consumers may be charged a price that is below that of perfect competition, as long as the marginal revenue exceeds the marginal cost.

In **second-degree price discrimination** the monopolist offers a menu of quantity-based pricing options designed to induce customers to self-select based on how highly they value the product. Such mechanisms include volume discounts, volume surcharges, coupons, product bundling, and restrictions on use. In practice, producers can use not just the quantity but also the quality (e.g., professional grade) to charge more to customers who value the product highly.

Third-degree price discrimination happens when customers are segregated by demographic or other traits. For example, some econometric software is licensed this way: A student version can handle only small data sets and is sold for a low price; a professional version can handle very large data sets and is sold at a much higher price because corporations need to compute the estimates for their business and are therefore willing to pay more for a license. Another example is that airlines know that passengers who want to fly somewhere and come back the same day are most likely businesspeople; therefore, one-day round-trip tickets are generally more expensive than tickets with a return flight at a later date or over a weekend.

Price discrimination has many practical applications when the seller has pricing power. The best way to understand how this concept works is to think of consumer surplus: As seen in this chapter, a consumer may be willing to pay more for the first unit of a good, but to buy a second unit she will want to pay a lower price, thus getting a better deal on the first unit. In practice, sellers can sometimes use income and substitution effects to their advantage. Think of something you often buy, perhaps lunch at your favorite café. How much would you be willing to pay for a lunch club membership card that would allow you to purchase lunches at, say, half price? If the café could extract from you the maximum amount each month that you would be willing to pay for the half-price option, then it would successfully have removed the income effect from you in the form of a monthly fixed fee. Notice that a downward-sloping demand curve implies that you would end up buying more lunches each month than before you purchased the discount card, even though you would be no better or worse off than before. This is a way that sellers are sometimes able to extract consumer surplus by means of creative pricing schemes. It's a common practice among big-box retailers, sports clubs, and other users of what is called two-part tariff pricing, as in Example 4-3.

EXAMPLE 4-3 Price Discrimination

Nicole's monthly demand for visits to her health club is given by the following equation: $Q_D = 20 - 4P$, where Q_D is visits per month and P is euros per visit. The health club's marginal cost is fixed at €2 per visit.

1. Draw Nicole's demand curve for health club visits per month.
2. If the club charged a price per visit equal to its marginal cost, how many visits would Nicole make per month?
3. How much consumer surplus would Nicole enjoy at that price?
4. How much could the club charge Nicole each month for a membership fee?

Solution to 1: $Q_D = 20 - 4P$, so when $P = 0$, $Q_D = 20$. Inverting, $P = 5 - 0.25\ Q_D$, so when $Q = 0$, $P = 5$.

Solution to 2: $Q_D = 20 - 4(2) = 12$. Nicole would make 12 visits per month at a price of €2 per visit.

Solution to 3: Nicole's consumer surplus can be measured as the area under her demand curve and above the price she pays for a total of 12 visits, or $(0.5)(12)(3) = 18$. Nicole would enjoy a consumer surplus of €18 per month.

Solution to 4: The club could extract all of Nicole's consumer surplus by charging her a monthly membership fee of €18 plus a per-visit price of €2. This pricing method is called a two-part tariff because it assesses one price per unit of the item purchased plus a per-month fee (sometimes called an entry fee) equal to the buyer's consumer surplus evaluated at the per-unit price.

6.5. Factors Affecting Long-Run Equilibrium in Monopoly Markets

The unregulated monopoly market structure can produce economic profits in the long run. In the long run all factors of production are variable, whereas in the short run some factors of production are fixed. Generally, the short-run factor that is fixed is the capital investment, such as the factory, machinery, production technology, available arable land, and so forth. The long-run solution allows for all inputs, including technology, to change. In order to maintain a monopoly market position in the long run, the firm must be protected by substantial and ongoing barriers to entry. If the monopoly position is the result of a patent, then new patents must be continuously added to prevent the entry of other firms into the market.

For regulated monopolies, such as natural monopolies, there are a variety of long-run solutions. One solution is to set the price equal to marginal cost, $P = \text{MC}$. However, that price will not likely be high enough to cover the average cost of production, as Exhibit 4-19 illustrated. The answer is to provide a subsidy sufficient to compensate the firm. The national rail system in the United States, Amtrak, is an example of a regulated monopoly operating with a government subsidy.

National ownership of the monopoly is another solution. Nationalization of the natural monopoly has been a popular solution in Europe and other parts of the world. The United States has generally avoided this potential solution. One problem with this arrangement is that once a price is established, consumers are unwilling to accept price increases, even as factor costs increase. Politically, raising prices on products from government-owned enterprises is highly unpopular.

Establishing a governmental entity that regulates an authorized monopoly is another popular solution. Exhibit 4-19 illustrated the appropriate decision rule. The regulator sets

price equal to long-run average cost, P_R = LRAC. This solution ensures that investors will receive a normal return for the risk they are taking in the market. Given that no other competitors are allowed, the risk is lower than in a highly competitive market environment. The challenge facing the regulator is determining the authentic risk-related return and the monopolist's realistic long-run average cost.

The final solution is to franchise the monopolistic firm through a bidding war. Again, the public goal is to select the winning firm based on price equaling long-run average cost. Retail outlets at rail stations and airports and concession outlets at stadiums are examples of government franchises. The long-run success of the monopoly franchise depends on its ability to meet the goal of pricing its products at the level of its long-run average cost.

EXAMPLE 4-10 Monopolies and Efficiency

Are monopolies *always* inefficient?

A. No, because if they charge more than average cost they can be nationalized.
B. Yes, because they charge all consumers more than perfectly competitive markets would.
C. No, because economies of scale and regulation (or threat of entry) may give a better outcome for buyers than perfect competition might provide.

Solution: C is correct. Economies of scale and regulation may make monopolies more efficient than perfect competition.

7. IDENTIFICATION OF MARKET STRUCTURE

Monopoly markets and other situations where companies have pricing power can be inefficient because producers constrain output to cause an increase in prices. Therefore, there will be less of the good being consumed and it will be sold at a higher price, which is generally inefficient for the market as a whole. This is why competition law regulates the degree of market competition in several industries of different countries.

Market power in the real world is not always as clear as it is in textbook examples. Governments and regulators often have the difficult task of measuring market power and establishing whether a firm has a dominant position that may resemble a monopoly. For example, in the 1990s, U.S. regulators prosecuted agricultural corporation Archer Daniels Midland (ADM) for conspiring with Japanese competitors to fix the price of lysine, an amino acid used as an animal feed additive. The antitrust action resulted in a settlement that involved over US$100 million in fines paid by the cartel members. Another example occurred in the 1970s, when U.S. antitrust authorities broke up the local telephone monopoly, leaving AT&T the long-distance business (and opening that business to competitors), and required AT&T to divest itself of the local telephone companies it owned. This antitrust decision brought competition, innovation, and lower prices to the U.S. telephone market.

European regulators (specifically, the European Commission) have affected the mergers and monopoly positions of European corporations (as in the case of the companies Roche,

Rhone-Poulenc, and BASF, which were at the center of a vitamin price-fixing case) as well as non-European companies (such as Intel) that do business in Europe. Moreover, the merger between the U.S. company General Electric and the European company Honeywell was denied by the European Commission on grounds of excessive market concentration.

Quantifying excessive market concentration is difficult. Sometimes, regulators need to measure whether something that has not yet occurred might generate excessive market power. For example, a merger between two companies might allow the combined company to be a monopolist or quasi-monopolist in a certain market.

A financial analyst hearing news about a possible merger should always consider the impact of competition law (sometimes called antitrust law)—that is, whether a proposed merger may be blocked by regulators in the interest of preserving a competitive market.

7.1. Econometric Approaches

How should one measure market power? The theoretical answer is to estimate the elasticity of demand and supply in a market. If demand is very elastic, the market must be very close to perfect competition. If demand is rigid (inelastic), companies *may* have market power. This is the approach taken in the cellophane case mentioned in Section 3.1.2.

From the econometric point of view, this estimation requires some attention. The problem is that observed price and quantity are the equilibrium values of price and quantity and do not represent the value of either supply or demand. Technically, this is called the problem of endogeneity, in the sense that the equilibrium price and quantity are jointly determined by the interaction of demand and supply. Therefore, to have an appropriate estimation of demand and supply, we will need to use a model with two equations, namely, an equation of demanded quantity (as a function of price, income of the buyers, and other variables) and an equation of supplied quantity (as a function of price, production costs, and other variables). The estimated parameters will then allow us to compute elasticity.

Regression analysis is useful in computing elasticity but requires a large number of observations. Therefore, one may use a time-series approach and, for example, look at 20 years of quarterly sales data for a market. However, the market structure may have changed radically over those 20 years, and the estimated elasticity may not apply to the current situation. Moreover, the supply curve may change as a result of a merger among large competitors, and the estimation based on past data may not be informative regarding the future state of the market after the merger.

An alternative approach is a cross-sectional regression analysis. Instead of looking at total sales and average prices in a market over time (the time-series approach mentioned earlier), we can look at sales from different companies in the market during the same year, or even at single transactions from many buyers and companies. Clearly, this approach requires substantial data-gathering effort, and therefore this estimation method can be complicated. Moreover, different specifications of the explanatory variables—for example, using total gross domestic product (GDP) rather than median household income or per-capita GDP to represent income—may sometimes lead to dramatically different estimates.

7.2. Simpler Measures

Trying to avoid the aforementioned drawbacks, analysts often use simpler measures to estimate elasticity. The simplest measure is the concentration ratio, which is the percentage of market share held by the N largest firms in an industry. To compute this ratio, one would, for

example, add the sales values of the 10 largest firms and divide this figure by total market sales. This number is always between zero (perfect competition) and 100 percent (monopoly).

The main advantage of the concentration ratio is that it is simple to compute, as just shown. The disadvantage is that it does not directly quantify market power. In other words, is a high concentration ratio a clear signal of monopoly power? The analysis of entry in Section 2 explains clearly that this is not the case: A company may be the only incumbent in a market, but if the barriers to entry are low, the simple presence of a *potential* entrant may be sufficient to convince the incumbent to behave like a firm in perfect competition. For example, a sugar wholesaler may be the only one in a country, but the knowledge that other large wholesalers in the food industry might easily add imported sugar to their range of products should convince the sugar wholesaler to price its product as if it were in perfect competition.

Another disadvantage of the concentration ratio is that it tends to be unaffected by mergers among the top market incumbents. For example, if the largest and second-largest incumbents merge, the pricing power of the combined entity is likely to be larger than that of the two preexisting companies. But the concentration ratio may not change much.

Calculating the Concentration Ratio

Suppose there are eight producers of a certain good in a market. The largest producer has 35 percent of the market, the second largest has 25 percent, the third has 20 percent, the fourth has 10 percent, and the remaining four have 2.5 percent each. If we computed the concentration ratio of the top three producers, it would be $35 + 25 + 20 = 80$ percent, while the concentration ratio of the top four producers would be $35 + 25 + 20 + 10 = 90$ percent.

If the two largest companies merged, the new concentration ratio for the top three producers would be 60 (the sum of the market shares of the merged companies) $+ 20 + 10 = 90$ percent, and the concentration ratio for the four top producers would be 92.5 percent. Therefore, this merger affects the concentration ratio very mildly, even though it creates a substantial entity that controls 60 percent of the market.

For example, the effect of consolidation in the U.S. retail gasoline market has resulted in increasing degrees of concentration. In 1992, the top four companies in the U.S. retail gasoline market shared 33 percent of the market. By 2001, the top four companies controlled 78 percent of the market (Exxon Mobil 24 percent, Shell 20 percent, BP/Amoco/Arco 18 percent, and Chevron/Texaco 16 percent).

To avoid the known issues with concentration ratios, economists O. C. Herfindahl and A. O. Hirschman suggested an index where the market shares of the top N companies are first squared and then added. If one firm controls the whole market (a monopoly), the Herfindahl-Hirschman index (HHI) equals 1. If there are M firms in the industry with equal market shares, then the HHI equals $1/M$. This provides a useful gauge for interpreting an HHI. For example, an HHI of 0.20 would be analogous to having the market shared equally by five firms.

The HHI for the top three companies in the calculation example in the box would be $0.35^2 + 0.25^2 + 0.20^2 = 0.225$ before the merger, whereas after the merger it would be $0.60^2 + 0.20^2 + 0.10^2 = 0.410$, which is substantially higher than the initial 0.225. This is why the HHI is widely used by competition regulators. However, just like the concentration ratio, the HHI does not take the possibility of entry into account, nor does it consider the elasticity of demand. As a consequence, the HHI has limited use for a financial analyst trying to estimate the potential profitability of a company or group of companies.

EXAMPLE 4-12 The Herfindahl-Hirschman Index

Suppose a market has 10 suppliers, each of them with 10 percent of the market. What are the concentration ratio and the HHI of the top four firms?

A. Concentration ratio 4 percent and HHI 40
B. Concentration ratio 40 percent and HHI 0.4
C. Concentration ratio 40 percent and HHI 0.04

Solution: C is correct. The concentration ratio for the top four firms is $10 + 10 + 10 + 10 = 40$ percent, and the HHI is $0.10^2 \times 4 = 0.01 \times 4 = 0.04$.

8. SUMMARY

In this chapter, we have surveyed how economists classify market structures. We have analyzed the distinctions among the different structures that are important for understanding demand and supply relations, optimal price and output, and the factors affecting long-run profitability. We also provided guidelines for identifying market structure in practice. Among our conclusions are the following:

- Economic market structures can be grouped into four categories: perfect competition, monopolistic competition, oligopoly, and monopoly.
- The categories differ because of the following characteristics: The number of producers is many in perfect and monopolistic competition, few in oligopoly, and one in monopoly. The degree of product differentiation, the pricing power of the producer, the barriers to entry of new producers, and the level of nonprice competition (e.g., advertising) are all low in perfect competition, moderate in monopolistic competition, high in oligopoly, and generally highest in monopoly.
- A financial analyst must understand the characteristics of market structures in order to better forecast a firm's future profit stream.
- The optimal marginal revenue equals marginal cost. However, only in perfect competition does the marginal revenue equal price. In the remaining structures, price generally exceeds marginal revenue because a firm can sell more units only by reducing the per-unit price.
- The quantity sold is highest in perfect competition. The price in perfect competition is usually lowest, but this depends on factors such as demand elasticity and increasing returns

to scale (which may reduce the producer's marginal cost). Monopolists, oligopolists, and producers in monopolistic competition attempt to differentiate their products so they can charge higher prices.
- Typically, monopolists sell a smaller quantity at a higher price. Investors may benefit from being shareholders of monopolistic firms that have large margins and substantial positive cash flows.
- Competitive firms do not earn economic profit. There will be a market compensation for the rental of capital and of management services, but the lack of pricing power implies that there will be no extra margins.
- While in the short run firms in any market structure can have economic profits, the more competitive a market is and the lower the barriers to entry, the faster the extra profits will fade. In the long run, new entrants shrink margins and push the least efficient firms out of the market.
- Oligopoly is characterized by the importance of strategic behavior. Firms can change the price, quantity, quality, and advertisement of the product to gain an advantage over their competitors. Several types of equilibrium (e.g., Nash, Cournot, kinked demand curve) may occur that affect the likelihood of each of the incumbents (and potential entrants in the long run) having economic profits. Price wars may be started to force weaker competitors to abandon the market.
- Measuring market power is complicated. Ideally, econometric estimates of the elasticity of demand and supply should be computed. However, because of the lack of reliable data and the fact that elasticity changes over time (so that past data may not apply to the current situation), regulators and economists often use simpler measures. The concentration ratio is simple, but the HHI, with little more computation required, often produces a better figure for decision making.

PRACTICE PROBLEMS[19]

1. A market structure characterized by many sellers with each having some pricing power and product differentiation is *best* described as:
   A. oligopoly.
   B. perfect competition.
   C. monopolistic competition.

2. A market structure with relatively few sellers of a homogeneous or standardized product is *best* described as:
   A. oligopoly.
   B. monopoly.
   C. perfect competition.

3. Market competitors are *least likely* to use advertising as a tool of differentiation in an industry structure identified as:
   A. monopoly.
   B. perfect competition.
   C. monopolistic competition.

[19]These practice problems were written by Tim Mahoney, CFA (Greenville, Rhode Island, USA).

4. Upsilon Natural Gas, Inc. is a monopoly enjoying very high barriers to entry. Its marginal cost is $40 and its average cost is $70. A recent market study has determined that the price elasticity of demand is 1.5. The company will *most likely* set its price at:
   A. $40.
   B. $70.
   C. $120.

5. The demand schedule in a perfectly competitive market is given by $P = 93 - 1.5Q$ (for $Q \leq 62$) and the long-run cost structure of each company is:

Total cost:	$256 + 2Q + 4Q^2$
Average cost:	$256/Q + 2 + 4Q$
Marginal cost:	$2 + 8Q$

   New companies will enter the market at any price greater than:
   A. 8.
   B. 66.
   C. 81.

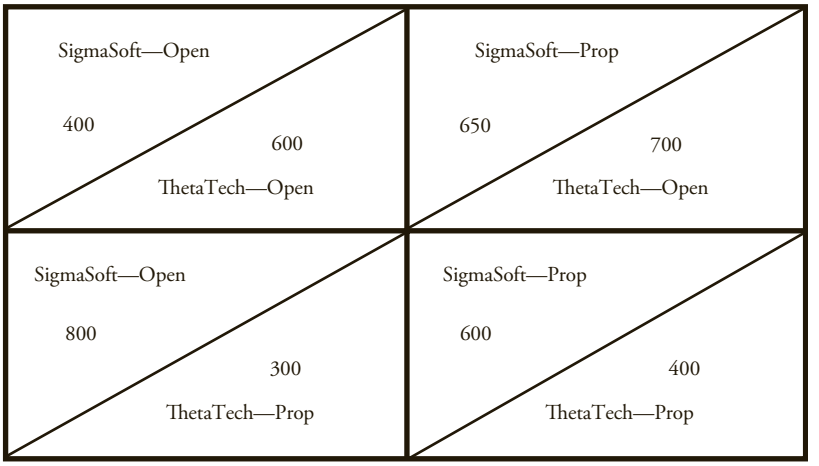
6. Companies *most likely* have a well-defined supply function when the market structure is:
   A. oligopoly.
   B. perfect competition.
   C. monopolistic competition.

7. Aquarius, Inc. is the dominant company and the price leader in its market. One of the other companies in the market attempts to gain market share by undercutting the price set by Aquarius. The market share of Aquarius will *most likely*:
   A. increase.
   B. decrease.
   C. stay the same.

8. SigmaSoft and ThetaTech are the dominant makers of computer system software. The market has two components: a large mass-market component in which demand is price sensitive, and a smaller performance-oriented component in which demand is much less price sensitive. SigmaSoft's product is considered to be technically superior. Each company can choose one of two strategies:

   1. *Open architecture (Open):* Mass-market focus allowing other software vendors to develop products for its platform.
   2. *Proprietary (Prop):* Allowing only its own software applications to run on its platform.

Depending on the strategy each company selects, their profits would be:



The Nash equilibrium for these companies is:
A. proprietary for SigmaSoft and proprietary for ThetaTech.
B. open architecture for SigmaSoft and proprietary for ThetaTech.
C. proprietary for SigmaSoft and open architecture for ThetaTech.

9. A company doing business in a monopolistically competitive market will *most likely* maximize profits when its output quantity is set such that:
   A. average cost is minimized.
   B. marginal revenue equals average cost.
   C. marginal revenue equals marginal cost.

10. Oligopolistic pricing strategy *most likely* results in a demand curve that is:
    A. kinked.
    B. vertical.
    C. horizontal.

11. Collusion is *less likely* in a market when:
    A. the product is homogeneous.
    B. companies have similar market shares.
    C. the cost structures of companies are similar.

12. If companies earn economic profits in a perfectly competitive market, over the long run the supply curve will *most likely*:
    A. shift to the left.
    B. shift to the right.
    C. remain unchanged.

13. Over time, the market share of the dominant company in an oligopolistic market will *most likely*:
    A. increase.
    B. decrease.
    C. remain the same.

14. A government entity that regulates an authorized monopoly will *most likely* base regulated prices on:
    A. marginal cost.
    B. long-run average cost.
    C. first-degree price discrimination.

15. An analyst gathers the following market share data for an industry:

Company	Sales (in millions of €)
ABC	300
Brown	250
Coral	200
Delta	150
Erie	100
All others	50

    The industry's four-company concentration ratio is *closest* to:
    A. 71 percent.
    B. 86 percent.
    C. 95 percent.

16. An analyst gathers the following market share data for an industry comprised of five companies:

Company	Market Share (%)
Zeta	35
Yusef	25
Xenon	20
Waters	10
Vlastos	10

The industry's three-firm Herfindahl-Hirschman index is *closest* to:
A. 0.185.
B. 0.225.
C. 0.235.

17. One disadvantage of the Herfindahl-Hirschman index is that the index:
A. is difficult to compute.
B. fails to reflect low barriers to entry.
C. fails to reflect the effect of mergers in the industry.

18. In an industry comprised of three companies that are small-scale manufacturers of an easily replicable product unprotected by brand recognition or patents, the *most* representative model of company behavior is:
A. oligopoly.
B. perfect competition.
C. monopolistic competition.

19. Deep River Manufacturing is one of many companies in an industry making a food product. Deep River units are identical up to the point they are labeled. Deep River produces its labeled brand, which sells for $2.20 per unit, and house brands for seven different grocery chains that sell for $2.00 per unit. Each grocery chain sells both the Deep River brand and its house brand. The *best* characterization of Deep River's market is:
A. oligopoly.
B. perfect competition.
C. monopolistic competition.

CHAPTER 5

AGGREGATE OUTPUT, PRICES, AND ECONOMIC GROWTH

Paul R. Kutasovic, CFA

Richard G. Fritz

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Calculate and explain gross domestic product (GDP) using expenditure and income approaches.
- Compare the sum-of-value-added and value-of-final-output methods of calculating GDP.
- Compare nominal and real GDP, and calculate and interpret the GDP deflator.
- Compare GDP, national income, personal income, and personal disposable income.
- Explain the fundamental relationship among saving, investment, the fiscal balance, and the trade balance.
- Explain the investment–saving (IS) and liquidity preference–money supply (LM) curves and how they combine to generate the aggregate demand curve.
- Explain the aggregate supply curve in the short run and the long run.
- Explain the causes of movements along and shifts in the aggregate demand and supply curves.
- Describe how fluctuations in aggregate demand and aggregate supply cause short-run changes in the economy and the business cycle.
- Explain how a short-run macroeconomic equilibrium may occur at a level above or below full employment.
- Analyze the effect of combined changes in aggregate supply and demand on the economy.
- Describe the sources, measurement, and sustainability of economic growth.
- Describe the production function approach to analyzing the sources of economic growth.
- Distinguish between input growth and growth of total factor productivity as components of economic growth.

1. INTRODUCTION

In the field of economics, *microeconomics* is the study of the economic activity and behavior of individual economic units, such as a household, a company, or a market for a particular good or service, and *macroeconomics* is the study of the aggregate activities of households, companies, and markets. Macroeconomics focuses on national aggregates, such as total *investment*, the amount spent by all businesses on plant and equipment; total *consumption*, the amount spent by all households on goods and services; the rate of change in the general level of prices; and the overall level of interest rates.

Macroeconomic analysis examines a nation's aggregate output and income, its competitive and comparative advantages, the productivity of its labor force, its price level and inflation rate, and the actions of its national government and central bank. The objective of macroeconomic analysis is to address such fundamental questions as:

- What is an economy's aggregate output, and how is aggregate income measured?
- What factors determine the level of aggregate output/income for an economy?
- What are the levels of aggregate demand and aggregate supply of goods and services within the country?
- Is the level of output increasing or decreasing, and at what rate?
- Is the general price level stable, rising, or falling?
- Is unemployment rising or falling?
- Are households spending or saving more?
- Are workers able to produce more output for a given level of inputs?
- Are businesses investing in and expanding their productive capacity?
- Are exports and imports rising or falling?

From an investment perspective, investors must be able to evaluate a country's current economic environment and to forecast its future economic environment in order to identify asset classes and securities that will benefit from economic trends occurring within that country. Macroeconomic variables—such as the level of inflation, unemployment, consumption, government spending, and investment—affect the overall level of activity within a country. They also have different impacts on the growth and profitability of industries within a country, the companies within those industries, and the returns of the securities issued by those companies.

This chapter is organized as follows: Section 2 describes gross domestic product and related measures of domestic output and income. Section 3 discusses short-run and long-run aggregate demand and supply curves; the causes of shifts and movements along those curves; and factors that affect equilibrium levels of output, prices, and interest rates. Section 4 discusses sources, sustainability, and measures of economic growth. A summary and practice problems complete the chapter.

2. AGGREGATE OUTPUT AND INCOME

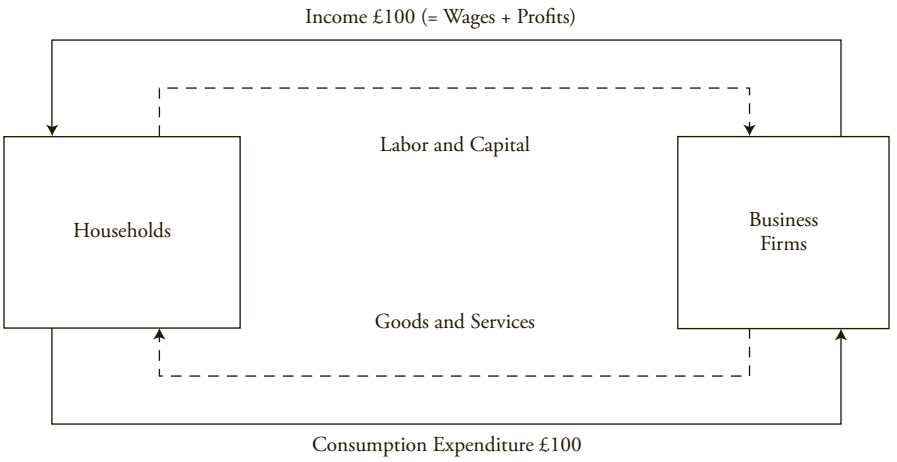
The **aggregate output** of an economy is the value of all the goods and services produced in a specified period of time. The **aggregate income** of an economy is the value of all the payments earned by the suppliers of factors used in the production of goods and services. Because the value of the output produced must accrue to the factors of production, aggregate output and aggregate income within an economy must be equal.

There are four broad forms of payments (i.e., income): compensation of employees, rent, interest, and profit. Compensation of employees includes wages and benefits (primarily employer contributions to private pension plans and health insurance) that individuals receive in exchange for providing labor. **Rent** is payment for the use of property. **Interest** is payment for lending funds. **Profit** is the return that owners of a company receive for the use of their capital and the assumption of financial risk when making their investments. Although businesses are the direct owners of much of the property and physical capital in the economy by virtue of owning the businesses, households are the ultimate owners of these assets and hence the ultimate recipients of the profits. In reality, of course, a portion of profits is usually retained within businesses to help finance maintenance and expansion of capacity. Similarly, because the government is viewed as operating on a nonprofit basis, any revenue it receives from ownership of companies or property may be viewed as being passed back to households in the form of lower taxes. Therefore, for simplicity, it is standard in macroeconomics to attribute all income to the household sector unless the analysis depends on a more precise accounting.

Aggregate *expenditure*, the total amount spent on the goods and services produced in the (domestic) economy during the period, must also be equal to aggregate output and aggregate income. However, some of this expenditure may come from foreigners in the form of net exports.[1] Thus, aggregate output, aggregate income, and aggregate expenditure all refer to different ways of decomposing the same quantity.

Exhibit 5-1 illustrates the flow of inputs, outputs, income, and expenditures in a very simple economy. Households supply the factors of production (labor and capital) to businesses in exchange for wages and profit (aggregate income) totaling £100. These flows are shown by the top two arrows. Companies use the inputs to produce goods and services (aggregate output), which they sell to households (aggregate expenditure) for £100. The output and

EXHIBIT 5-1 Output, Income, and Expenditure in a Simple Economy: The Circular Flow



[1]Note that *aggregate expenditure* as defined here does not equal the amount spent by domestic residents on goods and services, because it includes exports (purchases of domestic products by foreigners) and excludes imports (purchases of foreign products by domestic residents). Thus, spending by domestic residents does not necessarily equal domestic income/output. Indeed, within any given period, it usually does not. This will be explained in more detail in Section 2.2.3.

expenditure flows are shown by the bottom two arrows. Aggregate output, income, and expenditure are all equal to £100.

In this simplified example, households spend all of their income on domestically produced goods and services. They do not buy foreign goods, save for the future, or pay taxes. Similarly, businesses do not sell to foreigners or to the government and do not invest to increase their productive capacity. These important components of the economy will be added in Section 2.2. But first we need to discuss how output and income are measured.

2.1. Gross Domestic Product

Gross domestic product (GDP) measures:

- The market value of all final goods and services produced within the economy in a given period of time (output definition) or, equivalently,
- The aggregate income earned by all households, all companies, and the government within the economy in a given period of time (income definition).

Intuitively, GDP measures the flow of output and income in the economy.[2] GDP represents the broadest measure of the value of economic activity occurring within a country during a given period of time.

Therefore, GDP can be determined in two different manners. In the income approach, GDP is calculated as the total amount earned by households and companies in the economy. In the expenditure approach, GDP is calculated as the total amount spent on the goods and services produced within the economy during a given period. For the economy as a whole, total income must equal total expenditures, so the two approaches yield the same result.

Many developed countries use a standardized methodology for measuring GDP. This methodology is described in the official handbook of the Organization for Economic Cooperation and Development (Paris: OECD Publishing). The OECD reports the national accounts for many developed nations. In the United States, the National Income and Product Accounts (NIPA; also called the national accounts, for short) is the official U.S. government accounting of all the income and expenditure flows in the U.S. economy. The national accounts are the responsibility of the U.S. Department of Commerce and are published in its *Survey of Current Business*. In Canada, similar data are available from Statistics Canada, whereas in China, the National Bureau of Statistics of China provides GDP data.

To ensure that GDP is measured consistently over time and across countries, the following three broad criteria are used:

1. All goods and services included in the calculation of GDP *must be produced during the measurement period*. Therefore, items produced in previous periods—such as houses, cars, machinery, or equipment—are excluded. In addition, transfer payments from the government sector to individuals, such as unemployment compensation or welfare benefits,

[2]Some textbooks and countries measure flows of income and output by using gross national product (GNP) rather than GDP. The difference is subtle but can be important in some contexts. GDP includes production within national borders regardless of whether the factors of production (labor, capital, and property) are owned domestically or by foreigners. In contrast, GNP measures output produced by domestically owned factors of production regardless of whether the production occurs domestically or overseas.

are excluded. Capital gains that accrue to individuals when their assets appreciate in value are also excluded.

2. The only goods and services included in the calculation of GDP are those whose value *can be determined by being sold in the market*. This enables the prices of goods or services to be objectively determined. For example, a liter of extra virgin olive oil is more valuable than a liter of spring water because the market price of extra virgin olive oil is higher than the market price of spring water. The value of labor used in activities that are not sold on the market, such as commuting or gardening, is also excluded from GDP. By-products of production processes are also excluded if they have no explicit market value, such as air pollution, water pollution, and acid rain.
3. Only the market value of final goods and services is included in GDP. Final goods and services are those that are not resold. *Intermediate goods* are goods that are resold or used to produce another good.[3] The value of intermediate goods is excluded from GDP because additional value is added during the production process, and all the value added during the entire production process is reflected in the final sale price of the finished good. An alternative approach to measuring GDP is summing all the value added during the production and distribution processes. The most direct approach, however, is to sum the market value of all the final goods and services produced within the economy in a given time period.

Two distinct, but closely related, measurement methods can be used to calculate GDP based on expenditures: value of final output and sum of value added. These two methods are illustrated in Exhibit 5-2. In this example, a farmer sells wheat to a miller. The miller grinds

EXHIBIT 5-2 Value of Final Product Equals Income Created

	Receipts at Each Stage (€)	Value Added (= Income Created) at Each Stage (€)	
Receipts of farmer from miller	0.15	0.15	Value added by farmer
Receipts of miller from baker	0.46	0.31	Value added by miller
Receipts of baker from retailer	0.78	0.32	Value added by baker
Receipts of retailer from final customer	1.00	0.22	Value added by retailer
	1.00	1.00	
	Value of final output	Total value added = Total income created	

[3] *Final goods* should not be confused with so-called final sales, and *intermediate goods* should not be confused with inventories. GDP includes both final sales to customers and increases in companies' inventories. If sales exceed current production, then GDP is less than final sales by the amount of goods sold out of inventory.

the wheat into flour and sells it to a baker, who makes bread and sells it to a retailer. Finally, the bread is sold to retail customers. The wheat and flour are both intermediate goods in this example because they are used as inputs to produce another good. Thus, they are not counted (directly) in GDP. For the purposes of GDP, the value of the final product is €1.00, which includes the value added by the bread retailer as a distributor of the bread. If, in contrast, the baker sold directly to the public, the value counted in GDP would be the price at which the baker sold the bread, €0.78. The left column of the exhibit shows the total revenue received at each stage of the process, whereas the right column shows the value added at each stage. Note that the market value of the final product (€1.00) is equal to the sum of the value added at each of the stages. Thus, the contribution to GDP can be measured as either the final sale price or the sum of the value added at each stage.

EXAMPLE 5-1 Contribution of Automobile Production to GDP

Exhibit 5-3 provides simplified information on the cost of producing an automobile in the United States at various stages of the production process. The example assumes the automobile is produced and sold domestically and assumes no imported material is used. Calculate the contribution of automobile production to GDP using the value-added method, and show that it is equivalent to the expenditure method. What impact would the use of imported steel or plastics have on GDP?

EXHIBIT 5-3 Cost of Producing Automobiles

Stage of Production	Sales Value ($)
1. Production of basic materials	
Steel	1,000
Plastics	3,000
Semiconductors	1,000
2. Assembly of automobile (manufacturer price)	15,000
3. Wholesale price for automobile dealer	16,000
4. Retail price	18,000

Solution: GDP includes only the value of final goods and ignores intermediate goods in order to avoid double counting. Thus, the final retail sale price of $18,000 and not the total of $54,000 from summing sales at all the levels of production would be included in GDP. Alternatively, we can avoid double counting by calculating and summing the value added at each stage. At each stage of production, the difference between what a

company pays for its inputs and what it then receives for the product is its contribution to GDP. The value added for each stage of production is computed as follows:

Stage of Production	Sales Value ($)	Value Added ($)	
1. Production of basic materials			
Steel	1,000	1,000	
Plastics	3,000	3,000	
Semiconductors	1,000	1,000	
Total inputs		5,000	(sum of three inputs)
2. Assembly of car (manufacturer price)	15,000	10,000	= (15,000 − 5,000)
3. Wholesale price for car dealer	16,000	1,000	= (16,000 − 15,000)
4. Retail price	18,000	2,000	= (18,000 − 16,000)
Total expenditure	18,000		
Total value added		18,000	

Thus, the sum of the value added by each stage of production is equal to $18,000, which is equal to the final selling price of the automobile. If some of the inputs (steel, plastics, or semiconductors) are imported, the value added would be reduced by the amount paid for the imports.

2.1.1. Goods and Services Included at Imputed Values

As a general rule, only the value of goods and services whose *value can be determined by being sold in the market* are included in the measurement of GDP. Owner-occupied housing and government services, however, are two examples of services that are not sold in the marketplace but are still included in the measurement of GDP.

When a household or individual rents a place to live, the renter is buying housing services. The household pays the owner of the property rent in exchange for shelter. The income that a property owner receives is included in the calculation of GDP. However, when a household purchases a home, it is implicitly paying itself in exchange for the shelter. As a result, the government must estimate (impute) a value for this owner-occupied rent, which is then added to GDP.

The value of government services provided by police officers, firefighters, judges, and other government officials is a key factor that affects the level of economic activity. However, valuing these services is difficult because they are not sold in a market like other services; individual customers cannot decide how much to consume or how much they are willing to pay. Therefore, these services are simply included in GDP at their cost (e.g., wages paid) with no value added attributed to the production process.

For simplicity and global comparability, the number of goods and services with imputed values that are included in the measurement of GDP is limited. In general, nonmarket activity

is excluded from GDP. Thus, activities performed for one's own benefit, such as cooking, cleaning, and home repair, are excluded. Activities in the so-called underground economy are also excluded. The underground economy reflects economic activity that people hide from the government either because it is illegal or because they are attempting to evade taxation. Undocumented laborers who are paid off the books are one example. The illegal drug trade is another. Similarly, barter transactions, such as neighbors exchanging services with each other (for example, helping your neighbor repair her fence in exchange for her plowing your garden), are excluded from GDP.

Exhibit 5-4 shows the estimated size of the underground economy in various countries as a percentage of nominal GDP. The estimates range from 8 percent in the United States to 60 percent in Peru.

It should be clear from these estimates of the underground economy that the reliability of official GDP data varies considerably across countries. Failure to capture a significant portion of activity is one problem. Poor data collection practices and unreliable statistical methods within the official accounts are also potential problems.

EXHIBIT 5-4 Underground Economy as a Percentage of Nominal GDP, 2006

Country	Underground Economy as a Percentage of Nominal GDP (%)
Peru	60.0
Mexico	32.1
South Korea	27.5
Costa Rica	26.8
Greece	26.0
India	24.4
Italy	23.1
Spain	20.2
Sweden	16.3
Germany	15.4
Canada	14.1
China	14.0
France	13.2
Japan	8.9
United States	8.0

Source: Friedrich Schneider and Andreas Buehn, "Shadow Economies and Corruption All Over the World: Revised Estimates for 120 Countries," *University of Linz, Austria; Technische Universität Dresden, Germany*, Vol. 1, 2007–9, July 24, 2007. www.economics-ejournal.org/economics/journalarticles/2007-9. Version 2, October 27, 2009.

2.1.2. Nominal and Real GDP

In order to evaluate an economy's health, it is often useful to remove the effect on GDP of changes in the general price level, because higher (or lower) income driven solely by changes in the price level is not indicative of a higher (or lower) level of economic activity. To accomplish this, economists use **real GDP**, which indicates what would have been the total expenditures on the output of goods and services if prices were unchanged. **Per capita real GDP** (real GDP divided by the size of the population) has often been used as a measure of the average standard of living in a country.

Suppose we are interested in measuring the GDP of an economy. For the sake of simplicity, suppose that the economy consists of a single automobile maker and that in 2011, 300,000 vehicles are produced with an average market price of €18,750. GDP in 2011 would be €5,625,000,000. Economists define the value of goods and services measured at current prices as **nominal GDP**. Suppose that in 2012, 300,000 vehicles are again produced but that the average market price for a vehicle increases by 7 percent to €20,062.50. GDP in 2012 would be €6,018,750,000. Even though no more cars were produced in 2012 than in 2011, it appears that the economy grew by 7 percent between 2011 and 2012 (€6,018,750,000/€5,625,000,000) − 1 = 7%), although it actually did not grow at all.

Nominal and real GDP can be expressed as:

$$\text{Nominal GDP}_t = P_t \times Q_t$$

where:

P_t = Prices in year t

Q_t = Quantity produced in year t

$$\text{Real GDP}_t = P_B \times Q_t$$

where:

P_B = Prices in the base year

Taking the base year to be 2011 and putting in the 2011 and 2012 numbers gives:

$$\begin{aligned}
\text{Nominal GDP}_{2011} &= (€18{,}750 \times 300{,}000) = €5{,}625{,}000{,}000 \\
\text{Real GDP}_{2011} &= (€18{,}750 \times 300{,}000) = €5{,}625{,}000{,}000 \\
\text{Nominal GDP}_{2011} &= (€20{,}062.50 \times 300{,}000) = €6{,}018{,}750{,}000 \\
\text{Real GDP}_{2012} &= (€18{,}750 \times 300{,}000) = €5{,}625{,}000{,}000
\end{aligned}$$

In this example, real GDP did not change between 2011 and 2012, because the total output remained the same: 300,000 vehicles. The difference between nominal GDP in 2012 and real GDP in 2012 was the 7 percent inflation rate.

Now suppose that the auto manufacturer produced 3 percent more vehicles in 2012 than in 2011 (i.e., production in 2012 was 309,000 vehicles). Real GDP would increase by 3 percent from 2011 to 2012. With a 7 percent increase in prices, nominal GDP for 2012 would now be:

$$\begin{aligned}
\text{Nominal GDP}_{2012} &= (1.03 \times 300{,}000) \times (1.07 \times €18{,}750) \\
&= 309{,}000 \times €20{,}062.50 \\
&= €6{,}199{,}312{,}500
\end{aligned}$$

The **implicit price deflator for GDP**, or simply the **GDP deflator**, is defined as:

$$\text{GDP deflator} = \frac{\text{Value of current-year output at current-year prices}}{\text{Value of current-year output at base-year prices}} \times 100$$

Thus, in the example the GDP deflator for 2012 is $[(309{,}000 \times €20{,}062.50)/(309{,}000 \times €18{,}750)](100) = (1.07)(100) = 107$. The GDP deflator broadly measures the aggregate changes in prices across the overall economy, and hence changes in the deflator provide a useful gauge of inflation within the economy.

The real GDP is equal to the nominal GDP divided by the GDP deflator scaled by 100:

$$\text{Real GDP} = \text{Nominal GDP}/(\text{GDP deflator}/100)$$

This relationship gives the GDP deflator its name. That is, the measure of GDP in terms of current prices, nominal GDP, is adjusted for inflation by dividing it by the deflator. The expression also shows that the GDP deflator is the ratio of nominal GDP to the real GDP scaled by 100:

$$\text{GDP deflator} = (\text{Nominal GDP}/\text{Real GDP}) \times 100$$

Thus, real GDP for 2012 would be:

$$\begin{aligned}\text{Real GDP}_{2012} &= \text{Nominal GDP}/(\text{GDP deflator}/100)\\ &= €6{,}199{,}312{,}500/(107/100)\\ &= €5{,}793{,}750{,}000\end{aligned}$$

Note that €5,793,750,000 represents 3 percent real growth over 2011 GDP and 3 percent higher real GDP for 2012 than under the assumption of no growth in unit car sales in 2012.

What would be the increase in *nominal* GDP for 2012 compared with 2011 with the 3 percent greater automobile production and 7 percent inflation?

$$\begin{aligned}&(\text{Nominal GDP}_{2012}/\text{Nominal GDP}_{2011}) - 1\\ &= (€6{,}199{,}312{,}500/€5{,}625{,}000{,}000) - 1\\ &= 0.102\end{aligned}$$

So, nominal GDP would increase by 10.2 percent, which equals $(1.07 \times 1.03) - 1$ or approximately $7\% + 3\% = 10\%$. Which number is more informative about growth in economic activity, 3 percent real growth or 10.2 percent nominal growth? The real growth rate is more informative because it exactly captures increases in output. Nominal growth, by blending price changes with output changes, is less directly informative about output changes. In summary, real economic growth is measured by the percentage change in real GDP. When measuring real economic activity or when comparing one nation's economy to another's, real GDP and real GDP growth should be used because they more closely reflect the quantity of output available for consumption and investment.

EXAMPLE 5-2 Calculating the GDP Deflator

John Lambert is an equity analyst with Equitytrust, a Canadian investment management firm that primarily invests in Canadian stocks and bonds. The investment policy committee for the firm is concerned about the possibility of inflation. The implicit GDP price deflator is an important measure of the overall price level in the economy, and changes in the deflator provide an important gauge of inflation within the economy. GDP data have been released by Statistics Canada and are shown in Exhibit 5-5. Lambert is asked by the committee to use the GDP data to calculate the implicit GDP price deflator from 2005 to 2009 and the inflation rate for 2009.

EXHIBIT 5-5 Real and Nominal GDP for Canada

Seasonally adjusted at annual rates (SAAR)					
	2005	2006	2007	2008	2009
GDP at market prices (million C$)	1,373,845	1,450,405	1,529,589	1,599,608	1,527,258
Real GDP (million 2002 C$)	1,247,807	1,283,033	1,311,260	1,318,054	1,285,604

Solution: The implicit GDP price deflator measures inflation across all sectors of the economy, including consumers, business, government, exports, and imports. It is calculated as the ratio of nominal to real GDP and reported as an index number with the base-year deflator equal to 100. The implicit GDP price deflator for the Canadian economy for 2009 is calculated as $(1{,}527{,}258/1{,}285{,}604) \times 100 = 118.8$. The results for 2009 and the other years are:

	2005	2006	2007	2008	2009
GDP at market prices (million C$)	1,373,845	1,450,405	1,529,589	1,599,608	1,527,258
Real GDP (million 2002 C$)	1,247,807	1,283,033	1,311,260	1,318,054	1,285,604
Implicit GDP price deflator	110.1	113.0	116.6	121.4	118.8

The inflation rate is calculated as a percentage change in the index. For 2009, the annual inflation rate is equal to $(118.8/121.4) - 1 = -2.1$ percent. This shows that Canada actually experienced deflation in 2009 even though prices are still above their level in 2007.

2.2. The Components of GDP

Having defined GDP and discussed how it is measured, we can now consider the major components of GDP, the flows among the four major sectors of the economy—the household sector, the business sector, the government sector, and the foreign or external sector (comprising transactions with the rest of the world)—and the markets through which they interact. An expression for GDP, based on the expenditure approach, is:

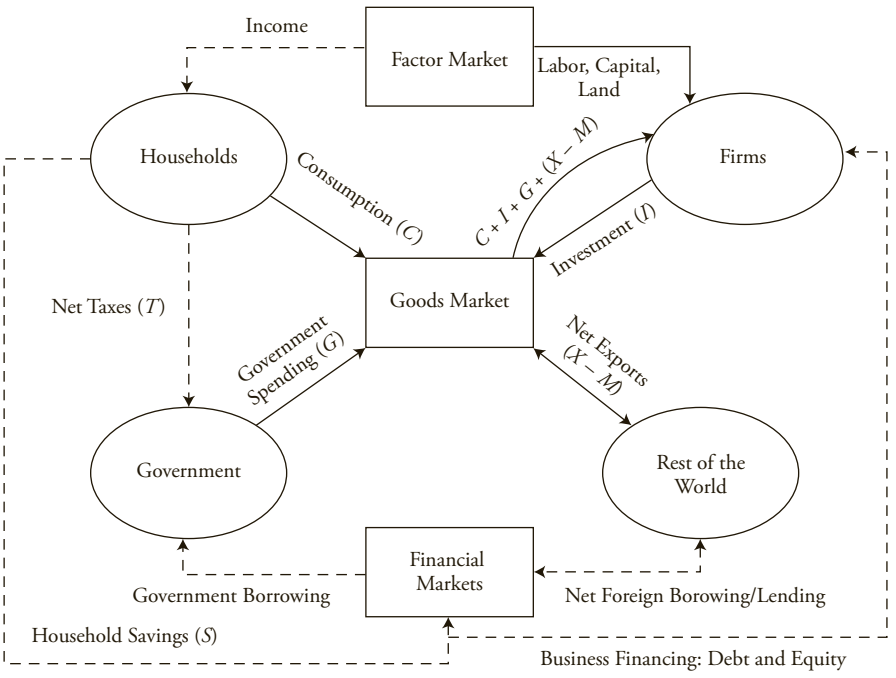
$$\text{GDP} = C + I + G + (X - M) \tag{5-1}$$

where:

C = Consumer spending on final goods and services
I = Gross private domestic investment, which includes business investment in capital goods (e.g., plant and equipment) and changes in inventory (**inventory investment**)
G = Government spending on final goods and services
X = Exports
M = Imports

Exhibit 5-6 shows the flow of expenditures, income, and financing among the four sectors of the economy and the three principal markets. In the exhibit, solid arrows point in the direction of expenditure on final goods and services. For simplicity, corresponding flows of output are not shown separately. The flow of factors of production is also shown with a solid

EXHIBIT 5-6 Output, Income, and Expenditure Flows



arrow. Financial flows, including income and net taxes, are shown with dashed arrows pointing to the recipient of funds.

2.2.1. The Household and Business Sectors

The very top portion of Exhibit 5-6 shows the services of labor, land, and capital flowing through the *factor market* to business firms and the flow of income back from firms to households. Households spend part of their income on consumption (*C*) and save (*S*) part of their income for future consumption. Current consumption expenditure flows through the *goods market* to the business sector. Household saving flows into the *financial markets*, where it provides funding for businesses that need to borrow or raise equity capital. Firms borrow or raise equity primarily to finance investment (*I*) in inventory, property, plant, and equipment. Investment (*I*) is shown flowing from firms through the goods market and back to firms because the business sector both demands and produces the goods needed to build productive capacity (*capital goods*).

In most developed economies, like Italy and the United States, expenditures on capital goods represent a significant portion of GDP. Investments (expenditures) on capital goods accounted for approximately 21.1 percent of Italy's GDP in 2007, while in the United States investments accounted for approximately 18.4 percent of GDP. In some developing countries, notably China (40.0 percent) and India (33.8 percent), investment spending accounts for a substantially larger share of the economy.[4] As we will examine in greater detail later, investment spending is an important determinant of an economy's long-term growth rate. At the same time, investment spending is the most volatile component of the economy, and changes in capital spending, especially spending on inventories, are one of the main factors causing short-run economic fluctuations.

2.2.2. The Government Sector

The government sector collects taxes from households and businesses. For simplicity, only the taxes collected from the household sector are shown in Exhibit 5-6. In turn, the government sector purchases goods and services (*G*) from the business sector. For example, the government sector hires construction companies to build roads, schools, and other infrastructure goods. Government expenditure (*G*) also reflects spending on the military, police and fire protection, the postal service, and other government services. Provision of these services makes the government a major source of employment in most countries. To keep Exhibit 5-6 simple, however, government employment and the corresponding income are not explicitly shown.

Governments also make transfer payments to households. In general, these are designed to address social objectives such as maintaining minimum living standards, providing health care, and assisting the unemployed with retraining and temporary support. In Exhibit 5-6, transfer payments are subtracted from taxes and reflected in net taxes (*T*).

Transfer payments are not included in government expenditures on goods and services (*G*) because they represent a monetary transfer by the government of tax revenue back to individuals with no corresponding receipt of goods or services. The household spending facilitated by the transfer payments is, of course, included in consumption (*C*) and, hence, in GDP. It is worth noting that transfers do not always take the form of direct payments to

[4]See Exhibit 5-27 later in this chapter for investment details for other countries. OECD.StatExtracts: Country Statistical Profiles 2009 (http://stats.oecd.org) and *Economic Report of the President* (Washington, DC: U.S. Government Printing Office, 2010): Table B-12, p. 345.

beneficiaries. Instead, the government may pay for or even directly provide goods or services to individuals. For example, universal health care programs often work in this way.

If, as is usually the case, government expenditure (G) exceeds net taxes (T), then the government has a *fiscal deficit* and must borrow in the financial markets. Thus, the government may compete with businesses in the financial markets for the funds generated by household saving. The only other potential source of funds in an economy is capital flows from the rest of the world. These will be discussed in the next section.

In 2007, the ratio of general government spending (which includes central government as well as state, provincial, and local government) to GDP in Italy was 44.8 percent while in the United States it was 31.2 percent. In countries where the government provides more services, such as universal health care in Italy, the government's contribution to GDP is greater. France's government sector represents 46.3 percent of GDP. In other countries, the public sector makes up a smaller share. For example, in Costa Rica, which has no standing army or navy, government spending is 16.1 percent of GDP. Exhibit 5-7 shows data on tax revenues, general government spending, and transfer payments as a share of nominal GDP as of 2007.

2.2.3. The External Sector

Trade and capital flows involving the rest of the world are shown in the bottom right quadrant of Exhibit 5-6. Net exports ($X - M$) reflects the difference between the value of goods and services sold to foreigners—exports (X)—and the portion of domestic consumption (C),

EXHIBIT 5-7 General Government Spending and Taxes as a Percentage of GDP, 2007

Country	General Government Tax Revenues as a Percentage of GDP	General Government Spending as a Percentage of GDP		
		Total	Goods and Services and Debt Service	Transfer Payments
Canada	33.3%	31.9%	17.1%	14.8%
Mexico	20.5	22.6	NA	NA
United States	28.3	31.2	19.5	11.7
Japan	28.1	30.5	14.9	15.6
South Korea	28.7	24.0	15.7	8.3
France	43.6	46.3	20.2	26.1
Germany	36.2	36.0	13.7	22.3
Greece	31.3	36.7	19.8	16.9
Italy	43.3	44.8	21.5	23.3
Spain	37.2	35.3	16.6	18.7
Sweden	48.2	44.5	17.7	26.8
Costa Rica	14.0	16.1	NA	NA

Sources: OECD.StatExtracts: Country Statistical Profiles 2009 (http://stats.oecd.org) and *Revenue Statistics 1965–2008: 2009 Edition* (Paris: OECD Publishing, 2009).

EXHIBIT 5-8 U.S. International Trade in Goods—Selected Countries, 2008 (millions of U.S. dollars)

	Exports	Imports	Balance
Total, all countries	1,276,994	1,508,109	–231,115
Europe	321,151	440,802	–119,651
Euro area	198,538	277,728	–79,190
France	28,603	44,036	–15,973
Germany	54,209	97,597	–43,388
Italy	15,330	36,140	–20,810
Canada	261,872	342,920	–81,048
Mexico	151,147	219,808	–68,661
China	69,552	337,963	–268,411
India	17,623	25,739	–8,116
Japan	64,457	139,587	–75,130

Source: Economic Report of the President (Washington, DC: U.S. Government Printing Office, 2010): Table B-105, p. 451.

investment (I), and government expenditure (G) that represents purchases of goods and services from the rest of the world—imports (M).

A **balance of trade deficit** means that the domestic economy is spending more on foreign goods and services than foreign economies are spending on domestic goods and services. It also means that the country is spending more than it produces because domestic saving is not sufficient to finance domestic investment plus the government's fiscal balance. A trade deficit must be funded by borrowing from the rest of the world through the financial markets. The rest of the world is able to provide this financing because, by definition, it must be running a corresponding trade surplus and spending less than it produces.

It bears emphasizing that trade and capital flows between an economy and the rest of the world must balance. One area's deficit is another's surplus, and vice versa. This is an accounting identity that must hold. In effect, having allowed a country to run a trade deficit, foreigners must, in aggregate, finance it. However, the financing terms may or may not be attractive.

Exhibit 5-8 reports trade balances for the United States with selected countries. Note that Canada was the largest trading partner, in terms of both exports and imports, in 2008. China was a close second in selling goods to U.S. markets, but China is not an important consumer of U.S. goods. Hence, the U.S. trade deficit with China was the largest by far. Overall, in 2008 the U.S. balance of trade deficit was $231,115 million.

2.3. GDP, National Income, Personal Income, and Personal Disposable Income

This section examines the calculation of GDP and other income measures in detail by means of an analysis of data from Statistics Canada.

Exhibit 5-9 provides data on the level of Canadian GDP and its components measured at market prices (nominal GDP), leaving certain quantities to be determined.

EXHIBIT 5-9 GDP Release for the Canadian Economy (millions of C$ at market prices, seasonally adjusted at annual rates)

	2005	2006	2007	2008	2009
Expenditure Based:					
Consumer spending	758,966	801,742	851,603	890,351	898,728
Government spending	259,857	277,608	293,608	314,329	333,942
Government gross fixed investment	37,067	41,151	45,321	50,955	59,078
Business gross fixed investment	255,596	283,382	301,885	313,574	269,394
Exports	519,435	524,075	534,718	563,948	438,553
Deduct: Imports	468,270	487,674	505,055	539,012	464,722
Change in inventories[a]	10,614	9,362	8,266	5,472	–8,180
Statistical discrepancy	580	759	–757	–9	465
GDP at Market Prices	**1,373,845**	**1,450,405**	**1,529,589**	**1,599,608**	**TBD**
Income Based:					
Wages, salaries, and supplementary labor income	695,093	743,392	784,885	818,613	819,066
Corporate profits before tax[b]	185,855	194,024	203,392	210,756	149,438
Government business enterprise profits before taxes	15,293	14,805	15,493	16,355	12,975
Interest income	61,421	66,404	71,589	83,998	63,947
Unincorporated business net income, including rent	85,234	86,750	90,411	94,559	99,879
Taxes less subsidies on factors of production	61,982	64,536	67,900	71,094	70,604
Taxes less subsidies on products	93,302	96,052	98,816	94,840	93,030
National Income	**1,198,180**	**1,265,963**	**1,332,486**	**1,390,215**	**TBD**
Statistical discrepancy	–581	–759	757	10	–466
Capital consumption allowance	176,246	185,201	196,346	209,383	218,785
GDP at Market Prices	**1,373,845**	**1,450,405**	**1,529,589**	**1,599,608**	**TBD**
Undistributed corporate profits	91,926	96,793	90,829	110,431	56,969
Corporate income taxes	51,631	47,504	54,867	53,176	34,319
Transfer payments: government to consumer	136,247	145,754	154,609	163,979	174,390
Personal Income	**1,035,586**	**1,106,832**	**1,174,683**	**1,224,653**	**TBD**
Personal disposable income	794,269	853,190	901,634	949,484	965,628
Interest paid to business	14,029	16,978	19,063	19,558	18,115

EXHIBIT 5-9 *Continued*

	2005	2006	2007	2008	2009
Consumer transfers to foreigners	4,395	4,483	5,533	5,117	4,737
Personal saving	16,878	29,987	25,435	34,458	TBD

[a]Includes change in government inventory.
[b]Includes inventory valuation adjustment.
Source: Statistics Canada.

The exhibit shows the two approaches to measuring GDP: (1) expenditures on final output measured as the sum of sales to the final users and (2) the sum of the factor incomes generated in the production of final output. In theory, the two approaches should provide the same estimate of GDP. As shown in the exhibit, however, in practice they differ because of the use of different data sources. The difference is accounted for by a *statistical discrepancy*. Market analysts more closely follow the expenditure approach because the expenditure data are more timely and reliable than data for the income components.[5]

Using the expenditure approach, Statistics Canada measures Canadian GDP as follows:

$$\begin{aligned}\text{GDP} = {} & \text{Consumer spending on goods and services} \\ & + \text{Business gross fixed investment} \\ & + \text{Change in inventories} \\ & + \text{Government spending on goods and services} \\ & + \text{Government gross fixed investment} \\ & + \text{Exports} - \text{Imports} \\ & + \text{Statistical discrepancy}\end{aligned}$$

Note that the Canadian national income accounts classify a portion of government expenditures as gross fixed investment. Not all countries make this distinction. The United States, for example, does not. Also note that the change in business inventories must be included in expenditures. Otherwise, goods produced but not yet sold would be left out of GDP.

The income-based approach calculates GDP as the sum of factor incomes and essentially measures the cost of producing final output. However, two of the costs entering into the gross value of output are not really earned by a factor of production. These items, depreciation and indirect taxes, are discussed later in this chapter. GDP is estimated in the income approach as follows:[6]

$$\begin{aligned}\text{GDP} = {} & \text{National income} + \text{Capital consumption allowance} \\ & + \text{Statistical discrepancy}\end{aligned}$$

[5]As shown in Exhibit 5-9, Statistics Canada divides the total statistical discrepancy roughly equally (with opposite signs) between the income-based and expenditure-based measures of GDP. In the U.S. national accounts, the statistical discrepancy appears only in the income-based breakdown of GDP because the expenditures data are believed to be more accurate than the income data.

[6]Construction of the national income accounts varies across countries. In the United States, for example, national income is defined to include income received by U.S.-owned factors of production even if the income is generated outside the country. To compute U.S. GDP, the national income data must be adjusted for net foreign factor income. No adjustment is required in the Canadian data since the data are measured on a geographic basis equivalent to GDP.

where **national income** is the income received by all factors of production used in the generation of final output:

National income = Compensation of employees
+ Corporate and government enterprise profits before taxes
+ Interest income
+ Unincorporated business net income (proprietor's income)
+ Rent
+ Indirect business taxes less subsidies

Compensation of employees includes wages and supplements to wages, which are primarily payments for pensions and health insurance. Corporate profits before taxes include three items: (1) dividends paid to households, (2) undistributed corporate profits (retained earnings) that remain in the business sector, and (3) corporate taxes paid to government. Interest income is the interest paid by businesses to households, government, and foreigners to compensate them for the loan of a financial asset. Unincorporated net income, including rent, is the earnings that flow to unincorporated proprietors and farm operators for running their own businesses. "Indirect business taxes less subsidies" reflects taxes and subsidies included in the final price of the good or service. It is the (net) portion of national income that is directly paid to the government. In the Canadian accounts, these are measured in two ways: (1) "taxes less subsidies on products," which include sales taxes, fuel taxes, and import duties, and (2) "taxes less subsidies on factors of production," which are mainly property taxes and payroll taxes.

The **capital consumption allowance** (CCA) is a measure of the wear and tear (depreciation) of the capital stock that occurs in the production of goods and services. This measure acknowledges the fact that some income/output must be allocated to replacement of the existing capital stock as it wears out. Loosely speaking, one may think of profit plus CCA as the total amount earned by capital, with the CCA being the amount that must be earned and reinvested just to maintain the existing productivity of the capital.

Along with the GDP report, Statistics Canada and other government statistical agencies provide information on personal income and saving. **Personal income** is a broad measure of household income and measures the ability of consumers to make purchases. As such, it is one of the key determinants of consumption spending. Personal income includes all income received by households, whether earned or unearned. It differs from national income in that some of the income earned by the factors of production (indirect business taxes, corporate income taxes, retained earnings) is not received by households and instead goes to the government or business sectors. Similarly, households receive some income from governments (transfer payments, such as social insurance payments, unemployment compensation, and disability payments) that is not earned. Thus, the following adjustments are made to national income in order to derive personal income:

Personal income = National income
− Indirect business taxes
− Corporate income taxes
− Undistributed corporate profits
+ Transfer payments

Personal disposable income (PDI) is equal to personal income less personal taxes. It measures the amount of after-tax income that households have to spend on goods and services or to save. Thus, it is the most relevant, and most closely watched, measure of income for household spending and saving decisions.

Finally, household saving is equal to PDI less three items: consumption expenditures, interest paid by consumers to business, and personal transfer payments to foreigners. The corresponding measure of saving for the business sector equals undistributed corporate profits plus the capital consumption allowance.

EXAMPLE 5-3 Canadian GDP Release and Other Measures of Production and Income

The investment policy committee at Equitytrust asks John Lambert to review the Canadian GDP data shown in Exhibit 5-9 and data from the Department of Finance Canada that show that the combined federal–provincial government deficit for 2009 was 84,249 (million C$), with the federal deficit at 55,590 (million C$).

1. Calculate 2009 GDP using the expenditure approach, and indicate how the expenditures are represented in Exhibit 5-6.
2. Calculate 2009 GDP using the income approach.
3. Calculate personal income for 2009.
4. Using the Canadian data for 2009, calculate the level of household saving (S), the saving rate, and net taxes (T) paid by the household sector. Given that the combined government budget deficit was 84,249 (million C$) in 2009, calculate tax revenues for the Canadian economy.
5. Calculate the impact of foreign trade on the Canadian economy in 2009 and Canada's net foreign borrowing/lending in 2009.
6. Calculate the net amount of borrowing/lending by the business sector in 2009.

Solutions (All numbers in millions):
Solution to 1: In the expenditure approach, nominal GDP is calculated as the sum of spending by the major sectors in the economy:

$$\begin{aligned}\text{GDP} = {} & \text{Consumer spending on goods and services} \\ & + \text{Business gross fixed investment} \\ & + \text{Change in inventories} \\ & + \text{Government spending on goods and services} \\ & + \text{Government gross fixed investment} \\ & + \text{Exports} - \text{Imports} \\ & + \text{Statistical discrepancy}\end{aligned}$$

Substituting the numbers from Exhibit 5-9,

$$\begin{aligned}\text{GDP} &= 898{,}728 + 269{,}394 + 333{,}942 + 59{,}078 + 438{,}553 \\ &\quad - 464{,}722 - 8{,}180 + 465 \\ &= \text{C\$}1{,}527{,}258\end{aligned}$$

In Exhibit 5-6, these expenditures are represented by the arrows pointing to the goods market and by the arrow pointing back to firms labeled as $C + I + G + (X - M)$.

Solution to 2: On the income side, nominal GDP is equal to national income plus the capital consumption allowance plus a statistical discrepancy. National income is defined as the sum of income received by the factors of production and is given by:

$$\begin{aligned}\text{National income} = {} & \text{Compensation of employees} \\ & + \text{Corporate and government enterprise profits before taxes} \\ & + \text{Interest income} \\ & + \text{Unincorporated business net income (proprietor's income)} \\ & + \text{Rent} + \text{Inventory valuation adjustment} \\ & + \text{Indirect business taxes less subsidies}\end{aligned}$$

Substituting in the numbers from Exhibit 5-9, we get C$1,308,939, where indirect business taxes are equal to 70,604 + 93,030 = C$163,634. Using this result,

$$\text{GDP} = 1{,}308{,}939 + 218{,}785 - 466 = \text{C\$}1{,}527{,}258$$

Solution to 3: Personal income is calculated as:

$$\begin{aligned}\text{Personal income} = {} & \text{National income} \\ & - \text{Indirect business taxes} \\ & - \text{Corporate income taxes} \\ & - \text{Undistributed corporate profits} \\ & + \text{Transfer payments}\end{aligned}$$

Substituting in the numbers from Exhibit 5-9,

$$\begin{aligned}\text{Personal income} &= 1{,}308{,}939 - (70{,}604 + 93{,}030) - 34{,}319 - 56{,}969 + 174{,}390 \\ &= \text{C\$}1{,}228{,}407\end{aligned}$$

Solution to 4: Household saving is equal to personal disposable income less three items: consumption expenditures, interest paid by consumers to business, and personal transfer payments to foreigners. Consumption (C) is given in Exhibit 5-9 as C$898,728. Substituting the numbers, saving (S) = 965,628 − 898,728 − 18,115 − 4,737 = C$44,048. The Canadian saving rate for 2009 equals (44,048/965,628) = 4.6%.

Net taxes paid by the household sector consists of two components: (1) taxes paid by households to the government minus (2) government transfer payments to households. From Exhibit 5-9, government transfer payments to households for 2009 were C$174,390. The tax outlay for households in 2009 is the difference between personal income and personal disposable income. Therefore, tax payments by households to government equal 1,228,407 − 965,628 = C$262,779.

Thus, net taxes going to government (T) from the household sector is C$88,389. However, personal taxes do not cover all sources of receipts for government. Government receipts also come from such sources as corporate income taxes, indirect taxes on businesses and consumers, and contributions for social insurance. The total tax receipts for all levels of government can be estimated from the deficit information. From Exhibit 5-9, government spending for 2009 totaled (333,942 + 59,078) = C$393,020. Therefore, total tax revenue from all sources is equal to 393,020 minus 84,249 or C$308,771.

Solution to 5: The international sector had a large impact on the Canadian economy in 2009. Exports declined sharply—by 22.2 percent—going from C$563,948 in 2008 to C$438,553 in 2009. Imports declined from C$539,012 in 2008 to C$464,722 in 2009, a 13.8 percent decrease. As a result, the Canadian economy moved from a trade surplus of C$24,936 in 2008 to a deficit of C$26,169 in 2009. This huge swing in the trade balance had a very significant negative impact on the Canadian economy and subtracted from GDP growth.

Canada funded the large trade deficit in 2009 by borrowing C$26,169 from the rest of the world through the financial markets. As discussed in Section 2.2.3, trade and capital flows between an economy and the rest of the world must balance. A trade deficit must be funded by a capital inflow.

Solution to 6: Borrowing by the business sector depends on the level of saving in the sector (i.e., internally generated funds) and the level of business investment in both fixed assets and inventories (i.e., the amount that must be financed). For 2009, gross saving in the business sector is equal to undistributed corporate profits (56,969) plus the capital consumption allowance (218,785). Thus, business saving is C$275,754. Because this number exceeds business fixed and inventory investment of C$261,214 (269,394 − 8,180), the business sector was a net lender of funds totaling C$14,540.

3. AGGREGATE DEMAND, AGGREGATE SUPPLY, AND EQUILIBRIUM

In this section, we build a model of aggregate demand and aggregate supply and use it to discuss how aggregate output and the level of prices are determined in the economy. **Aggregate demand** (AD) represents the quantity of goods and services that households, businesses, government, and foreign customers want to buy at any given level of prices. **Aggregate supply** (AS) represents the quantity of goods and services that producers are willing to supply at any given level of prices. It also reflects the amount of labor and capital that households are willing to offer into the marketplace at given real wage rates and cost of capital.

3.1. Aggregate Demand

As we will see, the aggregate demand curve looks like the ordinary demand curves that we encounter in microeconomics: quantity demanded increases as the price level declines. But our intuitive understanding of that relationship—a lower price allows us to buy more of a good *with a given level of income*—does not apply here, because income is not fixed. Instead,

aggregate income or expenditure is to be determined within the model along with the price level. Thus, we will need to explain the relationship between price and quantity demanded somewhat differently.

The **aggregate demand curve** represents the combinations of aggregate income and the price level at which two conditions are satisfied. First, aggregate expenditure equals aggregate income. As indicated in our discussion of GDP accounting, this must always be true after the fact. The new aspect here is the requirement that *planned* expenditure equals *actual* (or realized) income. To understand the distinction, consider business inventories. If businesses end up with more inventory than they planned, then the difference represents unplanned (or unintended) business investment, and actual output in the economy exceeded *planned* expenditure by that amount. Second, the available real money supply is willingly held by households and businesses.

The first condition—equality of planned expenditures and actual income/output—gives rise to what is called the *investment–saving (IS) curve*. The second condition—equilibrium in the money market—is embodied in what is called the *liquidity preference–money supply (LM) curve*. When we put them together, we get the aggregate demand curve.

3.1.1. Balancing Aggregate Income and Expenditure: The IS Curve

Total expenditure on domestically produced output comes from four sources: household consumption (C), investments (I), government spending (G), and net exports ($X-M$). This can be expressed as:

$$\text{Expenditure} = C + I + G + (X - M)$$

Personal disposable income is equal to GDP (Y) plus transfer payments (F) minus retained earnings and depreciation (= business saving, S_B) minus direct and indirect taxes (R). Households allocate disposable income between consumption of goods and services (C) and household saving (S_H). Therefore,

$$Y + F - S_B - R = C + S_H$$

Rearranging this equation, we get:

$$Y = C + S + T$$

where $T = (R - F)$ denotes net taxes and $S = (S_B + S_H)$ denotes total private sector saving.

Because total expenditures must be identical to aggregate income (Y), we have the following relationship:

$$C + S + T = C + I + G + (X - M)$$

By rearranging this equation, we get the following fundamental relationship among domestic saving, investment, the fiscal balance, and the trade balance:

$$S = I + (G - T) + (X - M) \tag{5-2}$$

This equation shows that domestic private saving is used or absorbed in one of three ways: investment spending (I), financing government deficits ($G - T$), and building up

financial claims against overseas economies provided there is a positive trade balance $[(X-M)>0]$. If there is a trade deficit $[(X-M)<0]$, then domestic private saving is being supplemented by inflows of foreign saving, and overseas economies are building up financial claims against the domestic economy.

By rearranging the identity, we can examine the implications of government deficits and surpluses:

$$G - T = (S - I) - (X - M)$$

A fiscal deficit $[(G-T)>0]$ implies that the private sector must save more than it invests $[(S-I)>0]$ or the country must run a trade deficit $[(X-M)<0]$ with corresponding inflow of foreign saving, or there may be a combination of both.

EXAMPLE 5-4 Foreign Capital Inflows Help Finance Government Deficits

The budgetary situation changed dramatically in Canada during 2009. As noted in Example 5-3, the Department of Finance Canada reported that in 2009 the combined federal–provincial government had a deficit of 84,249 (million C$). Thus, the government sector operated at a deficit that needed to be financed. How was this deficit financed?

Solution: Using the formula $G-T=(S-I)-(X-M)$ shows that a budget deficit is financed through either higher domestic saving (S), lower business investment (I), or borrowing from foreigners $(X-M)$. Private saving is given by:

$$\begin{aligned}\text{Private saving} = {} & \text{Household saving} + \text{Undistributed corporate profits} \\ & + \text{Capital consumption allowance}\end{aligned}$$

Household saving for 2009 is given in the solution to Question 4 of Example 5-3. Using that figure and the 2009 values for undistributed corporate profits and capital consumption allowance from Exhibit 5-9, we get:

$$\text{Private saving} = 44{,}048 + 56{,}969 + 218{,}785 = \text{C\$}319{,}802$$

Comparing this number to the level of private investment in 2009 shows that private sector saving exceeded investment spending by C$58,588: 319,802 − (269,394 − 8,180) = C$58,588. Thus, domestic private saving financed over 71 percent of the deficit (58,588/82,249).

To finance the rest of the government deficit, foreign imports (M) would have to exceed exports (X) by C$23,661. From Exhibit 5-9, the actual trade deficit (amount of foreign borrowing) was C$26,169, slightly greater than the amount required. This difference is largely due to the statistical discrepancy caused by different data sources being used for expenditure-based and income-based estimates of GDP.

Equation 5-2 is the key relationship that must hold in order for aggregate income and aggregate expenditure to be equal. Up to this point, we have treated it as simply an accounting identity. We now need to think of it as the outcome of explicit decisions on the part of households, businesses, government, and foreigners. When we do so, we are faced with the question of what underlies these decisions and how the requisite balance is established.

Economists have found that the dominant determinant of consumption spending is disposable income ($Y - S_B - T$). This can be expressed formally by indicating that consumption is a function $C(\cdot)$ of disposable income,

$$C = C(Y - S_B - T)$$

or, dropping the technically correct but practically insignificant adjustment for retained earnings and depreciation (S_B), a function of GDP minus net taxes,

$$C = C(Y - T)$$

When households receive an additional unit of income, some proportion of this additional income is spent and the remainder is saved. The **marginal propensity to consume** (MPC) represents the proportion of an additional unit of disposable income that is consumed or spent. Because the amount that is not spent is saved, the **marginal propensity to save** (MPS) is MPS = 1 − MPC.

According to the consumption function, either an increase in real income or a decrease in taxes will increase aggregate consumption. Somewhat more sophisticated models of consumption recognize that consumption depends not only on current disposable income but also on wealth. Except for the very rich, individuals tend to spend a higher fraction of their current income as their wealth increases because with higher current wealth, there is less need to save to provide for future consumption.

Exhibit 5-10 shows household consumption expenditures as a percentage of GDP for selected countries.

These figures reflect the *average propensity to consume* (APC)—that is, the ratio C/Y—rather than a measure of how the next unit of income would be divided between spending and saving, the MPC. However, they are reasonable proxies for the MPC in each country. Comparing Germany's 56.6 percent APC with Mexico's 65.4 percent, the implication is that

EXHIBIT 5-10 Household Final Consumption Expenditures as a Percentage of GDP, 2007

United States	70.1
Mexico	65.4
Italy	58.7
France	56.7
Germany	56.6
Canada	56.5
Japan	56.3

Source: OECD.StatExtracts: Country Statistical Profiles 2009 (http://stats.oecd.org).

the Mexican economy is more sensitive to changes in disposable household income than the German economy is. All other things being equal, macroeconomic policies that increase disposable household income, such as lowering government taxes, would have a larger impact on the economies of Mexico (65.4 percent) and the United States (70.1 percent) than similar policies would have in Germany (56.6 percent) or France (56.7 percent).

Companies are the primary source of investment spending (I). They make investment decisions in order to expand their stock of physical capital, such as building new factories or adding new equipment to existing facilities. A definition of physical capital is *any man-made aid to production*. Companies also buy investment goods, such as manufacturing plants and equipment to replace existing facilities and equipment that wear out. Total investment, including replacement of worn-out capital, is called *gross investment*, as opposed to *net investment*, which reflects only the addition of new capacity. GDP includes gross investment; hence the name *gross* domestic product. Total investment spending in such developed countries as Italy, Germany, the United Kingdom, and the United States ranged between 18 and 22 percent of GDP in 2007.[7]

Investment decisions depend primarily on two factors: the level of interest rates and aggregate output/income. The level of interest rates reflects the cost of financing investment. The level of aggregate output serves as a proxy for the expected profitability of new investments. When an economy is underutilizing its resources, interest rates are typically very low and yet investment spending often remains dormant because the expected return on new investments is also low. Conversely, when output is high and companies have little spare capacity, the expected return on new investments is high. Thus, investment decisions may be modeled as a decreasing function $I(\cdot, \cdot)$ of the **real interest rate** (nominal interest rate minus the expected rate of inflation) and an increasing function of the level of aggregate output. Formally,

$$I = I(r, Y)$$

where I is investment spending, r is the real interest rate, and Y is, as usual, aggregate income. This investment function leaves out some important drivers of investment decisions, such as the availability of new and better technology. Nonetheless, it reflects the two most important considerations: the cost of funding (represented by the real interest rate) and the expected profitability of the new capital (proxied by the level of aggregate output).

Many government spending decisions are insensitive to the current level of economic activity, the level of interest rates, the currency exchange rate, and other economic factors. Thus, economists often treat the level of government spending on goods and services (G) as an *exogenous* policy variable determined outside the macroeconomic model. In essence, this means that the adjustments required to maintain the balance among aggregate spending, income, and output must occur primarily within the private sector.

Tax policy may also be viewed as an exogenous policy tool. However, the actual amount of net taxes (T) collected is closely tied to the level of economic activity. Most countries impose income taxes or value-added taxes (VAT) or both that increase with the level of income or expenditure. Similarly, at least some transfer payments to the household sector are usually based on economic need and are hence inversely related to aggregate income. Each of

[7]OECD.StatExtracts: Country Statistical Profiles 2009 (http://stats.oecd.org) and *Economic Report of the President* (Washington, DC: U.S. Government Printing Office, 2010): Table B-2, p. 330. See Exhibit 5-27 later in this chapter for investment details on other countries.

these factors makes net taxes (T) rise and fall with aggregate income, Y. The government's fiscal balance can be represented as:

$$G - T = \overline{G} - t(Y)$$

where $\overline{G}$ is the exogenous level of government expenditure and $t(Y)$ indicates that net taxes are an (increasing) function of aggregate income, Y. The fiscal balance decreases (smaller deficit or larger surplus) as aggregate income (Y) increases, and increases as income declines. This effect is called an *automatic stabilizer* because it tends to mitigate changes in aggregate output.

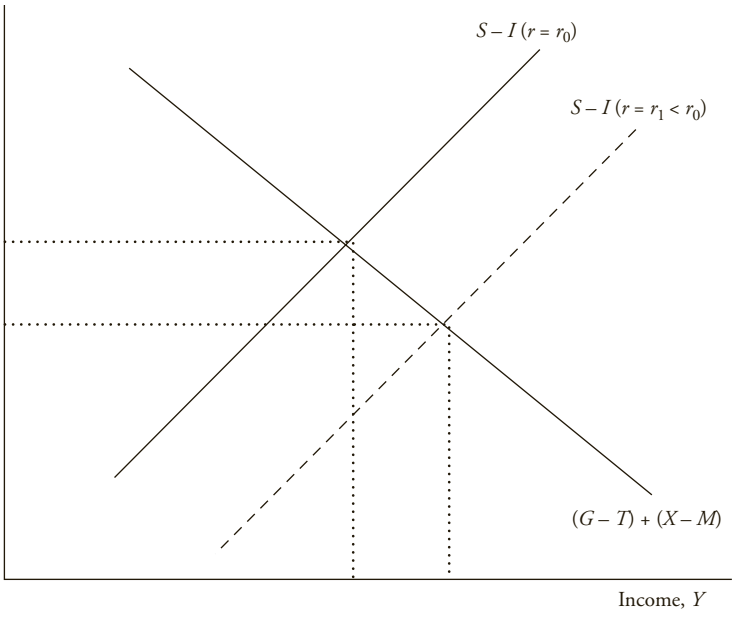
Net exports ($X - M$) are primarily a function of income in the domestic country and in the rest of the world along with the relative prices of domestic and foreign goods and services. As domestic income rises, some of the additional demand that is induced will be for imported goods. Thus, net exports will decline. An increase in income in the rest of the world will lead to an increase in demand for the domestic country's products and hence an increase net exports. A decrease in the relative prices of domestically produced goods and services, perhaps because of a depreciation of the currency, will shift demand toward these products and hence increase net exports.

We are now in a position to describe how aggregate expenditure and income are brought into balance. Slightly rearranging Equation 5-2, equality of expenditure and income implies:

$$S - I = (G - T) + (X - M)$$

Based on the previous discussion, we know that both the government's fiscal balance and the trade balance decrease as income rises because of net taxes and imports, respectively. Hence, the right-hand side of this equation declines with income. This is shown by the downward-sloping line in Exhibit 5-11. Assuming the direct effect of higher income on saving

EXHIBIT 5-11 Balancing Aggregate Income and Expenditure



is larger than the impact on investment, the left-hand side of the equation increases as income rises. This is shown by the solid upward-sloping line in Exhibit 5-11. Note that this line is drawn for a given level of the real interest rate, r_0. The intersection of these curves shows the level of income at which expenditure and income balance. At higher levels of income, the saving–investment differential $(S - I)$ exceeds the combined fiscal and trade balances, implying excess saving or insufficient expenditure. At lower levels of income, the saving–investment differential is smaller than the combined fiscal and trade balances, implying that planned expenditure exceeds output (= income).

The dashed, upward-sloping line in the exhibit reflects a lower real interest rate, $r_1 < r_0$. This line lies to the right of the solid line because for any value of the saving–investment differential $(S - I)$, the higher level of investment induced by a lower real interest rate requires a higher level of income to induce higher saving. With a lower real interest rate, the curves intersect at a higher level of income. Thus, we see that *equilibrating income and expenditure entails an inverse relationship between income and the real interest rate*. Economists refer to this relationship as the *IS curve* because investment (I) and saving (S) are the primary components that adjust to maintain the balance between aggregate expenditure and income. The IS curve is illustrated in Exhibit 5-12 and discussed in the next section.

EXAMPLE 5-5 The IS Curve

The following equations are given for a hypothetical economy:

$C = 2{,}000 + 0.7(Y - T)$	Consumption function
$I = 400 + 0.2Y - 30r$	Investment function
$G = 1{,}500$	Government spending
$(X - M) = 1{,}000 - 0.1Y$	Net export function
$T = -200 + 0.3Y$	Tax function

1. Based on these equations, determine the combinations of aggregate income (Y) and the real interest rate (r) that are consistent with equating income and expenditure. That is, find the equation that describes the IS curve.
2. Given a real interest rate of 4 percent, find the level of GDP, consumption spending, investment spending, net exports, and tax receipts.
3. Suppose the government increased expenditure from 1,500 to 2,000. Find the new IS curve. Does the increase in government spending result in an equal increase in equilibrium income for any given level of the real interest rate? Why or why not?
4. Given a real interest rate of 4 percent, determine how the increased government spending is funded.
5. Suppose that the output/income level calculated in Question 2 is the most that can be produced with the economy's resources. If the economy is operating at that level

when the government increases expenditure from 1,500 to 2,000, what must happen to maintain the balance between expenditure and income?

Solution to 1: Starting with the basic GDP identity $Y = C + I + G + (X - M)$ and substituting for each expenditure component using the equations above gives:

$$Y = 2{,}000 + 0.7(Y - T) + 400 + 0.2Y - 30r + 1{,}500 + 1{,}000 - 0.1Y$$

Substituting in the tax equation and solving for Y, we get:

$$\begin{aligned} Y &= 2{,}000 + 0.7(Y + 200 - 0.3Y) + 400 + 0.2Y - 30r + 1{,}500 + 1{,}000 - 0.1Y \\ &= 5{,}040 + 0.59Y - 30r \\ Y &= 12{,}292.7 - 73.2r \end{aligned}$$

The final equation is the IS curve. It summarizes combinations of income and the real interest rate at which income and expenditure are equal. Equivalently, it reflects equilibrium in the goods market.

Solution to 2: If the real interest rate is 4 percent, then GDP and the components of GDP are:

$$\begin{aligned} Y &= 12.292.7 - 73.2(4) = 11{,}999.9 \\ T &= -200 + 0.3(11{,}999.9) = 3{,}399.9 \\ C &= 2{,}000 + 0.7(11{,}999.9 - 3{,}399.9) = 8{,}020 \\ I &= 400 + 0.2(11{,}999.9) - 30(4) = 2{,}680.0 \\ (X - M) &= 1{,}000 - 0.10(11{,}999.9) = -200.0 \end{aligned}$$

Solution to 3: Following the same steps but with $G = 2{,}000$, the IS curve is:

$$Y = 13{,}512.2 - 73.2r$$

At any given level of the interest rate, aggregate income increases by $1{,}219.5 = (13{,}512.2 - 12{,}292.7)$. This is 2.44 (= 1,219.5/500) times the increase in government spending. The increase in government spending has a multiplier effect on equilibrium income because, as income rises, both consumption and investment spending also rise, leading to an even greater increase in income, which leads to even more spending. However, some of the increased private spending goes for imports, and higher income also induces higher taxes and saving. The condition for equality of income and expenditure can be written as:

$$G = (S - I) + T + (M - X)$$

So the increase in government spending must be balanced by some combination of (1) an increase in saving relative to investment, (2) an increase in taxes, and (3) a rise in

imports relative to exports. Given the interest rate, each of these will be induced by an increase in aggregate income. Because saving (S) equals $Y - C - T$,

$$\begin{aligned}\Delta S &= \Delta Y - \Delta C - \Delta T = \Delta Y - [0.7(\Delta Y - \Delta T)] - \Delta T \\ &= \Delta Y(1 - 0.7) + \Delta T(0.7 - 1) \\ &= 0.3\Delta Y - 0.3\Delta T = 0.3\Delta Y - 0.3(0.3)\Delta Y \\ &= 0.3(1 - 0.3)\Delta Y = 0.21\Delta Y\end{aligned}$$

Using this result along with the investment, tax, and trade balance functions gives:

$$\Delta G = (0.21 - 0.2)\Delta Y + 0.3\Delta Y + 0.1\Delta Y = 0.41\Delta Y$$

So, $\Delta Y = (1/0.41)\Delta G = 2.44\Delta G$.

Note that an extra unit of income increases saving by 0.21 but also increases investment spending by 0.20. So, in this hypothetical economy, the saving–investment differential ($S - I$) is very insensitive to the level of aggregate income. All else the same, this implies that relatively large changes in income are required to restore the expenditure/income balance whenever there is a change in spending behavior.

Solution to 4: Using the previous results,

$$\begin{aligned}\text{Change in fiscal balance} &= \Delta G - \Delta T = \Delta G[1 - 0.3(2.44)] \\ &= 0.268(500) = 134 \\ \text{Change in trade balance} &= \Delta(X - M) = 2.44\Delta G(-0.1) \\ &= -0.244(500) = -122 \\ \text{Change in } (S - I) &= \Delta(S - I) = 2.44\Delta G(0.21 - 0.20) \\ &= 0.0244(5) = 12\end{aligned}$$

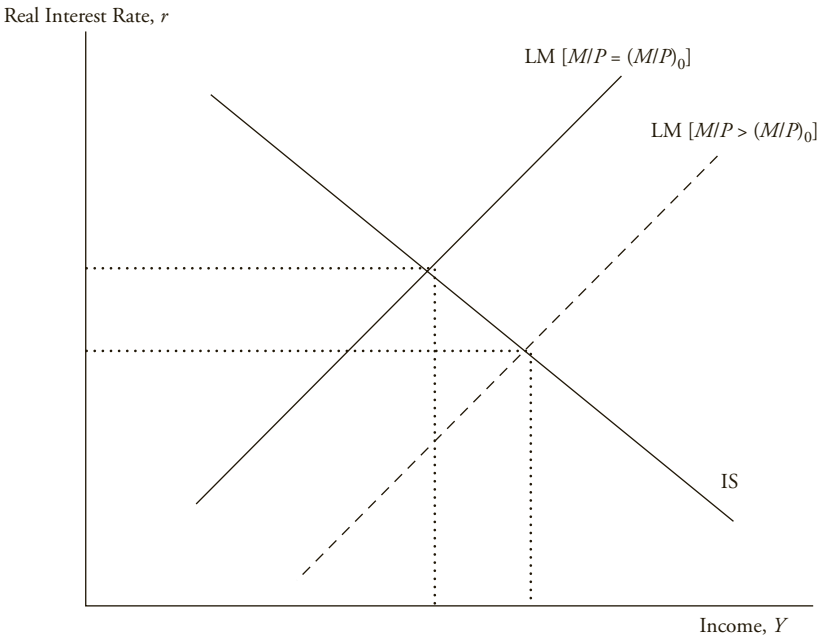
So, the increase in government spending (500) is ultimately financed by a large increase in taxes ($500 - 134 = 366$), a very small increase in private sector excess saving (12), and an increase in capital flows from abroad (122).

Solution to 5: If the economy is operating at maximum output, then an increase in government expenditure must crowd out an equal amount of private expenditure in order to keep total expenditure equal to output/income. In this simple model, this implies that the real interest rate must rise enough that investment spending falls by the amount of the increase in government spending. Using the new IS curve equation from the solution to Question 3 and the original level of income from Question 2, we need the interest rate such that:

$$Y = 13{,}512.2 - 73.2r = 11.999.9 \Rightarrow r = 20.66\%$$

So the real interest rate would soar from 4 percent to 20.66 percent to choke off investment spending.

EXHIBIT 5-12 The IS and LM Curves



3.1.2. Equilibrium in the Money Market: The LM Curve

The IS curve tells us what level of income is consistent with a given level of the real interest rate but does not address the appropriate level of interest rates, nor does it depend on the price level. In order to determine the interest rate and introduce a connection between output and the price level, we must consider supply and demand in the financial markets. To keep the model as simple as possible, we will deal explicitly with demand and supply for only one financial asset: money. All other assets (e.g., stocks and bonds) are implicitly treated as a composite alternative to holding money. In some of the subsequent discussion, however, we will note differential impacts on equity and fixed-income securities.

The *quantity theory of money* equation provides straightforward connections among the nominal money supply (M), the price level (P), and real income/expenditure (Y):

$$MV = PY$$

In this equation, V is the *velocity of money*, the average rate at which money circulates through the economy to facilitate expenditure. This equation essentially defines V. The equation begins to have economic content only when we make assumptions about how velocity is related to such economic variables as the interest rate. In the simplest case, if velocity is assumed to be constant, then the quantity theory of money equation implies that the money supply determines the nominal value of output (PY). Therefore, an increase in the money supply will increase the nominal value of output. However, this equation alone cannot tell us how that increase would be split between price and quantity.

The quantity theory equation can be rewritten in terms of the supply and demand for real money balances:

$$M/P = (M/P)_D = kY$$

where $k = 1/V$ and reflects how much money people want to hold for every currency unit of real income. The demand for real money balances is typically assumed to depend inversely on the interest rate because a higher interest rate encourages investors to shift their assets out of money (bank deposits) into higher-yielding securities. Although the quantity theory of money suggests that the demand for real money balances is proportional to real income, this need not be the case. The important point is that money demand increases with income. Thus, demand for real money balances is an increasing function $M(\cdot,\cdot)$ of real income and a decreasing function of the interest rate. Equilibrium in the money market requires:

$$M/P = M(r,Y)$$

Holding the real money supply (M/P) constant, this equation implies a positive relationship between real income (Y) and the real interest rate (r). Given the real money supply, an increase in real income must be accompanied by an increase in the interest rate in order to keep the demand for real money balances equal to the supply. This relationship, which economists refer to as the *liquidity preference–money supply (LM) curve*, is shown by the upward-sloping curve in Exhibit 5-12.

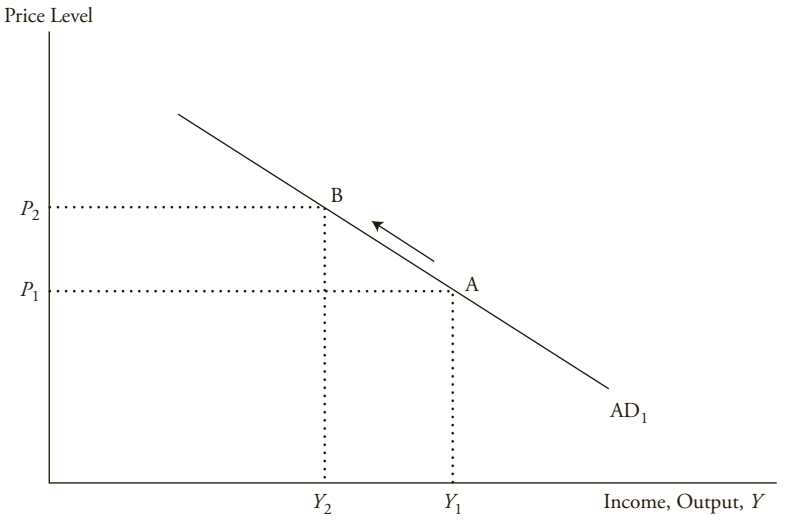
The intersection of the IS and LM curves determines the combination of real income and the real interest rate that is consistent with both the equality of income and (planned) expenditure (the IS curve) and equilibrium in the money market (the LM curve). In Exhibit 5-12, the dashed LM curve reflects a higher real money supply than the solid LM curve. With a higher real money supply, the intersection of the IS and LM curves occurs at a higher level of real income and a lower level of the real interest rate.

3.1.3. The Aggregate Demand Curve

If the nominal money supply (M) is held constant, then a higher or lower real money supply (M/P) arises because of changes in the lower price. If the price level declines, the real money supply increases and, as shown in Exhibit 5-12, real income increases while the real interest rate declines. Conversely, an increase in the price level leads to a decline in real income and an increase in the real interest rate. This inverse relationship between the price level and real income is illustrated in Exhibit 5-13. This is the **aggregate demand curve** (AD curve).

As shown in Exhibit 5-13, an increase in the price level from P_1 to P_2 reduces income from Y_1 to Y_2. Our development of the AD curve emphasizes only one channel through which prices affect the quantity of output demanded (i.e., planned real expenditure)—the interest rate. There are, however, other mechanisms. Higher prices erode the purchasing power of retirees and others whose income is fixed in nominal terms. Similarly, higher prices reduce the real value of nominal assets (e.g., stocks and bonds) and may reduce consumption relative to current income as people seek to rebuild the real purchasing power of their wealth. Higher domestic prices also make domestically produced goods more expensive relative to imports (assuming a constant currency exchange rate). In each case, lower prices have the opposite effect, increasing aggregate expenditure and income.

EXHIBIT 5-13 The Aggregate Demand Curve



It should be clear that many interesting and important aspects of the economy are subsumed into the AD curve: saving, investment, trade and capital flows, interest rates, asset prices, fiscal and monetary policy, and more. All of these disappear behind a deceptively simple relationship between price and output/income.

Before moving on to consider aggregate supply, let's look more closely at the interaction of interest rates and income implicit in movements along the AD curve. For simplicity, we assume there are no changes in the fiscal or trade balances so that maintaining the balance between aggregate expenditure and aggregate income requires that changes in investment spending equal changes in private saving. As the price level increases, the real money supply (M/P) declines. To induce a corresponding decline in money demand, the interest rate must rise so that other assets are more attractive, and income must fall to reduce the transactional need for money balances. The higher interest rate induces companies to reduce investment spending. The decline in income reduces household saving. *The slope of the AD curve depends on the relative sensitivities of investment, saving, and money demand to income and the interest rate*. The AD curve will be more flat if:

- Investment expenditure is highly sensitive to the interest rate.
- Saving is insensitive to income.
- Money demand is insensitive to interest rates.
- Money demand is insensitive to income.

The first two conditions directly imply that income will have to move more to induce a large enough change in saving to match the change in investment spending. All else being equal, each of the last two conditions implies that a larger change in the interest rate is required to bring money demand in line with money supply. This, in turn, implies a larger change in investment spending and a correspondingly larger change in saving and income.

EXAMPLE 5-6 Aggregate Demand

The money demand and supply equations for our hypothetical economy are:

$$M_d/P = -300 + 0.5Y - 30r \quad \text{(real money demand)}$$
$$M/P = 5{,}200/P \quad \text{(real money supply)}$$

1. Find the equation for the LM curve.
2. Using the IS curve from the solution to Question 1 of Example 5-5, find the equation of the AD curve.
3. Find the levels of GDP and the interest rate if $P = 1$.
4. What will happen to GDP and the interest rate if the price level rises to 1.1 or falls to 0.9?
5. Suppose investment spending were more sensitive to the interest rate so that the IS becomes ($Y = 12{,}292.7 - 150r$). What happens to the slope of the AD curve? What does this imply about the effectiveness of monetary policy?

Solution to 1: Setting the real money supply equal to the real money demand and rearranging, we get the LM equation:

$$Y = 600 + 2(M/P) + 60r$$

Or with $M = 5{,}200$,

$$Y = 600 + 10{,}400/P + 60r \quad \text{(LM equation)}$$

Solution to 2: From Question 1 of Example 5-5, the IS equation is $Y = 12{,}292.7 - 73.2\,r$. We now have two equations and two unknowns. The easiest way to solve this problem is to multiply the LM curve by 1.22 (= 73.2/60.0) and then add the equations:

$$1.22Y = 732 + 2.44(M/P) + 73.2r \quad \text{(LM equation)}$$
$$Y = 12{,}292.7 - 73.2r \quad \text{(IS equation)}$$

Adding the two equations and solving for Y,

$$Y = 5{,}867.0 + 1.099(M/P) \quad \text{(AD curve)}$$
$$= 5{,}867.0 + 5{,}715.3/P \quad (\text{with } M = 5{,}200)$$

Solution to 3: If $P = 1$, the AD curve gives GDP as $Y = 5{,}867.0 + 5{,}715.3 = 11{,}582.3$. From the money demand and supply equation, the equilibrium interest rate is:

$$5{,}200/1 = -300 + 0.5(11{,}582.3) - 30r \Rightarrow r = 9.7\%$$

Solution to 4: If the price level increases to 1.2, GDP declines to $Y = 5{,}867.0 + 5{,}715.3/1.1 = 11{,}062.7$. If the price level falls to 0.9, GDP increases to

$Y = 5{,}867.0 + 5{,}715.3/0.9 = 12{,}217.3$. To find the interest rate in each case, we plug these values for Y into the IS curve.

$$\text{If } P = 1.1\text{: } Y = 11{,}062.7 = 12{,}292.7 - 73.2\,r \Rightarrow r = 16.8\%$$
$$\text{If } P = 0.9\text{: } Y = 12{,}217.3 = 12{,}292.7 - 73.2\,r \Rightarrow r = 1.0\%$$

Thus, we have the following relationships among the price level, GDP, and the interest rate:

Price Level	GDP	Interest Rate
0.9	12,217.3	1.0
1.0	11,582.3	9.7
1.1	11,062.7	16.8

The inverse relationship between GDP and the price level is the AD curve. The inverse relationship between GDP and the interest rate reflects the IS curve.

Solution to 5: If the interest rate parameter in the IS curve is 150 instead of 73.2, we can multiply the LM equation by 2.5 (= 150/60) instead of 1.22 (= 73.2/60) to get the system of equations:

$$2.5Y = 1{,}500 + 5(M/P) + 150r \quad \text{(LM equation)}$$
$$Y = 12{,}292.7 - 150r \quad \text{(IS equation)}$$

Adding these equations and solving for Y gives:

$$Y = 3{,}940.77 + 1.429(M/P) \quad \text{(new AD curve)}$$
$$= 3{,}940.77 + 7{,}428.6/P \quad (\text{with } M = 5{,}200)$$

Comparing the new AD curve to the original AD curve indicates that output (Y) is now more sensitive to the price level. That is, the AD curve is more flat. Monetary policy is now more effective because, at any given price level, an increase in M has a greater impact on Y. This can be understood as follows: As the real money supply increases, the interest rate must fall and/or expenditure must increase in order to induce households to hold the increased money supply. With investment spending now more sensitive to the interest rate, income will have to rise by more in order to increase saving by a corresponding amount.

3.2. Aggregate Supply

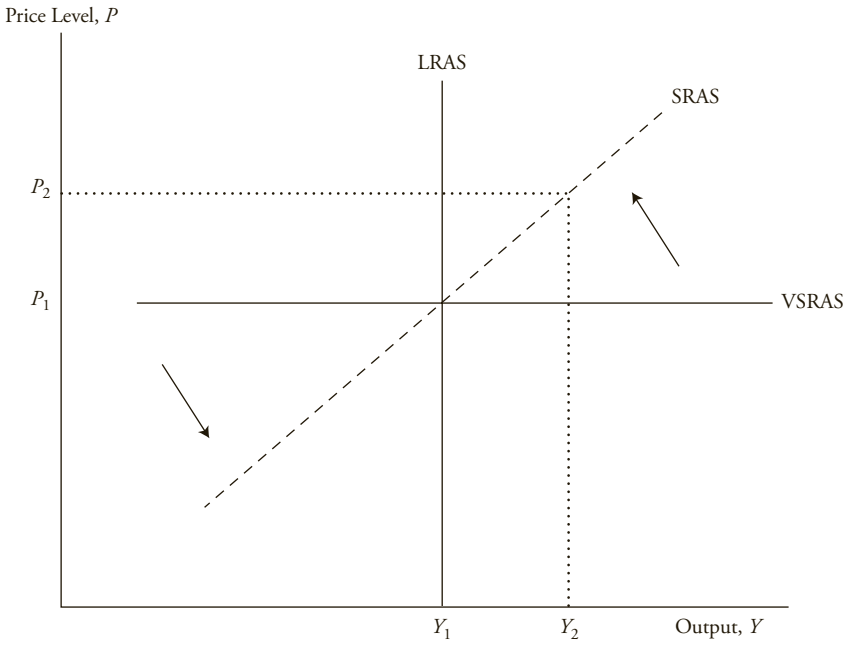
Aggregate demand only tells us the relationship between the price level and the amount of output demanded at those prices. To understand what price and output level will prevail in the economy, we need to add aggregate supply, the amount of output producers are willing to provide at various prices. The **aggregate supply curve** (AS curve) represents the level of

domestic output that companies will produce at each price level. Unlike the demand side, we must distinguish between the short- and long-run AS curves, which differ with respect to how wages and other input prices respond to changes in final output prices. *Long run* and *short run* are relative terms and are necessarily imprecise with respect to calendar time. The long run is long enough that wages, prices, and expectations can adjust but not long enough that physical capital is a variable input. Capital and the available technology to use that capital remain fixed. This condition implies a period of at least a few years and perhaps a decade. The truly long run in which even the capital stock is variable may be thought of as covering multiple decades. Consideration of the very long run is postponed to our discussion of economic growth in Section 4.

In the very short run, perhaps a few months or quarters, companies will increase or decrease output to some degree without changing price. This is shown in Exhibit 5-14 by the horizontal line labeled VSRAS (very short-run aggregate supply). If demand is somewhat stronger than expected, companies earn higher profit by increasing output as long as they can cover their variable costs. So they will run their plants and equipment more intensively, demand more effort from their salaried employees, and increase the hours of employees who are paid on the basis of hours worked. If demand is somewhat weaker than projected, companies can run their plants less intensively, cut labor hours, and utilize staff to perform maintenance and carry out efficiency-enhancing projects that are often postponed during busier periods.

Over somewhat longer periods, the AS curve is upward sloping because more costs become variable. This is represented by the short-run aggregate supply (SRAS) curve in Exhibit 5-14. In most businesses, wages are adjusted once a year, but for companies with union contracts, several years may pass before the contracts expire. The prices for raw materials

EXHIBIT 5-14 Aggregate Supply Curve



and other inputs may also be established under long-term contracts. Hence, wages and other input costs are relatively inflexible in the short run and do not fully adjust to changes in output prices. As the price level rises, most companies enjoy higher profit margins and hence expand production. In Exhibit 5-14, when prices move from P_1 to P_2, the quantity of aggregate output supplied increases from Y_1 to Y_2. Conversely, a reduction in the price level squeezes profit margins and causes companies to reduce production.

Over time, however, wages and other input prices tend to catch up with the prices of final goods and services. In other words, wages and prices that are inflexible or slow to adjust in the short run adjust to changes in the price level over the long run. Thus, over the long run, when the aggregate price level changes, wages and other input prices change proportionately so that the higher aggregate price level has no impact on aggregate supply. This is illustrated by the vertical long-run aggregate supply (LRAS) curve in Exhibit 5-14. As prices move from P_1 to P_2, the quantity of output supplied remains at Q_1 in the long run. The only change that occurs is that prices shift to a higher level (from P_1 to P_2).

The position of the LRAS curve is determined by the potential output of the economy. The amount of output produced depends on the fixed amount of capital and labor and the available technology. This classical model of aggregate supply can be expressed as:

$$Y = F(\overline{K}, \overline{L}) = \overline{Y}$$

where $\overline{K}$ is the fixed amount of capital and $\overline{L}$ is the available labor supply. The stock of capital is assumed to incorporate the existing technological base.[8] The available labor supply is also held constant, and workers are assumed to have a given set of skills. The long-run equilibrium level of output, Y_1 in Exhibit 5-14, is referred to as the *full employment*, or *natural*, level of output. At this level of output, the economy's resources are deemed to be fully employed and (labor) *unemployment is at its natural rate*. This concept of a natural rate of unemployment assumes the macroeconomy is currently operating at an efficient and unconstrained level of production. Companies have enough spare capacity to avoid bottlenecks, and there is a modest, stable pool of unemployed workers (job seekers equal job vacancies) looking for and transitioning into new jobs.

3.3. Shifts in Aggregate Demand and Supply

In the next two sections, the aggregate demand (AD) and aggregate supply (AS) models are used to address three critical macroeconomic questions:

1. What causes an economy to expand or contract?
2. What causes inflation and changes in the level of unemployment?
3. What determines an economy's rate of sustainable growth, and how can it be measured?

Before addressing these questions, we need to distinguish between (1) the long-run growth rate of real GDP and (2) short-run fluctuations in real GDP around this long-run trend.

The business cycle is a direct result of short-term fluctuations of real GDP. It consists of periods of economic expansion and contraction. In an expansion, real GDP is increasing, the

[8]Note that investment, I, reflects replacement of worn-out capital plus the change in capital, ΔK. Over short periods of time, net investment is assumed to have a negligible effect on aggregate supply. The cumulative effect of investment on economic growth is discussed in Section 4.

unemployment rate is declining, and capacity utilization is rising. In a contraction, real GDP is decreasing, the unemployment rate is rising, and capacity utilization is declining. Shifts in the AD and AS curves determine the short-run changes in the economy associated with the business cycle. In addition, the AD–AS model provides a framework for estimating the sustainable growth rate of an economy, which is addressed in Section 4.

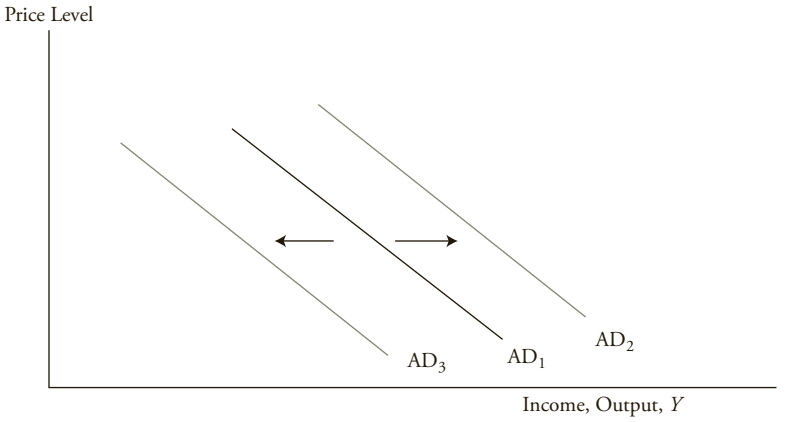
From an asset allocation perspective, it is important to determine the current phase of the business cycle as well as how fast the economy is growing relative to its sustainable growth rate. The expected rate of return on equities and fixed-income securities, for example, depends on estimates of the growth rate of GDP and inflation. For equities, GDP growth is the primary determinant of aggregate corporate profits. For fixed-income securities, the expected rate of inflation determines the spread between real and nominal rates of return. In order to use the AD and AS model to analyze the economy and to make investment decisions, we need to first understand what factors cause the curves to shift.

3.3.1. Shifts in Aggregate Demand

In addition to price, factors that influence the level of spending by households, companies, governments, and foreigners (i.e., the aggregate level of expenditures) will cause the AD curve to shift. A shift to the right represents an increase in aggregate demand at any price level. Exhibit 5-15 shows this as a shift from AD_1 to AD_2. A shift to the left represents a decrease in aggregate demand at any price level. This is indicated by a move from AD_1 to AD_3. Key factors that directly or indirectly influence the level of aggregate expenditures and cause the aggregate demand curve to shift include changes in:

- Household wealth.
- Consumer and business expectations.
- Capacity utilization.
- Monetary policy.
- The exchange rate.
- Growth in the global economy.
- Fiscal policy (government spending and taxes).

EXHIBIT 5-15 Shifts in the Aggregate Demand Curve



3.3.1.1. Household Wealth Household wealth includes the value of both financial assets (e.g., cash, savings accounts, investment securities, and pensions) and real assets (e.g., real estate). The primary reason households save a portion of their current income is to accumulate wealth for consumption in the future. The proportion of disposable income that households save depends partly on the value of the financial and real assets that they have already accumulated. If these assets increase in value, households will tend to save less and spend a greater proportion of their income because they will still be able to meet their wealth accumulation goals. As a result, an increase in household wealth increases consumer spending and shifts the aggregate demand curve to the right. In contrast, a decline in wealth will reduce consumer spending and shift the AD curve to the left. This is often referred to as the **wealth effect** and is one explanation for how changes in equity prices affect economic activity. Higher equity prices increase household wealth, which increases consumer spending and reduces the amount saved out of current income. Economic studies estimate that an increase or decrease in wealth in developed countries increases or decreases annual consumer spending by 3 to 7 percent of the change in wealth.[9] A smaller but still statistically significant wealth effect has been found in a number of emerging markets (developing countries).[10]

3.3.1.2. Consumer and Business Expectations Psychology has an important impact on consumer and business spending. When consumers are confident about their future income and the stability and safety of their jobs, they tend to spend a higher portion of their disposable income. This shifts the AD curve to the right. Consumer spending declines and the AD curve shifts to the left when consumers become less confident. Similarly, when businesses are optimistic about their future growth and profitability, they spend (invest) more on capital projects, which also shifts the AD curve to the right.

3.3.1.3. Capacity Utilization Capacity utilization is a measure of how fully an economy's production capacity is being used. Companies with excess capacity have little incentive to invest in new property, plant, and equipment. In contrast, when companies are operating at or near full capacity, they will need to increase investment spending in order to expand production. Data from the OECD and the U.S. Federal Reserve indicate that when aggregate capacity utilization reaches 82 to 85 percent, production blockages arise, prompting companies to increase their level of investment spending. This shifts the AD curve to the right.

3.3.1.4. Fiscal Policy **Fiscal policy** is the use of taxes and government spending to affect the level of aggregate expenditures.[11] An increase in government spending, one of the direct components of AD, shifts the AD curve to the right, whereas a decrease in government spending shifts the AD curve to the left. Taxes affect GDP indirectly through their effect on consumer spending and business investment. Lower taxes will increase the proportion of personal income and corporate pretax profits that consumers and businesses have available to spend and will shift the AD curve to the right. In contrast, higher taxes will shift the AD curve to the left.

[9]See, for example, Case, Quigley, and Shiller (2005).
[10]See Funke (2004).
[11]Government spending and taxes may be adjusted for other purposes, too. In macroeconomics, however, the term *fiscal policy* is usually reserved for actions intended to affect the overall level of expenditure.

EXAMPLE 5-7 The Wealth Effect on Saving and Consumption

The importance of the wealth effect on consumption, and its relationship to housing prices, was evident in the recession that began in late 2007. During this period, global GDP declined by the steepest amount in the post–World War II period. A major factor associated with the economic downturn was the sharp fall in housing prices, especially in countries that experienced a housing boom earlier in the decade, such as the United States, the United Kingdom, Spain, and Ireland. In each of these countries, consumers reduced spending sharply and raised the level of saving in response to the decline in wealth. Do the data in Exhibit 5-16 provide support for the wealth effect?

EXHIBIT 5-16 Housing Prices and the Saving Rate in the United Kingdom

Year	Housing Prices (first quarter of each year) (Index 2000 Q1 = 100)	Saving Rate (%)
2000	100	4.7
2002	122.7	5.8
2004	180.5	3.7
2006	206.3	2.9
2007	225.9	2.1
2008	220.5	1.2
2009	192.7	7.0

Source: Office of National Statistics, United Kingdom.

Solution: Housing prices in the United Kingdom rose by nearly 126 percent [(225.9 − 100)/100] between 2000 and 2007. As predicted, the saving rate declined (with a lag), going from an average of 5.3 percent of income in 2000 and 2002 to 1.2 percent in 2008. Then, as housing prices fell by 14.7 percent between 2007 and 2009, the saving rate rose dramatically from 1.2 percent in 2008 to 7 percent in 2009. Of course, the decline in housing prices was not the only factor contributing to the increase in the saving rate. Stock prices also declined in this period, further reducing wealth, and the recession raised uncertainty over future jobs and income.

3.3.1.5. Monetary Policy *Money* is generally defined as currency in circulation plus deposits at commercial banks. **Monetary policy** refers to action taken by a nation's central bank to affect aggregate output and prices through changes in bank reserves, reserve requirements, or the central bank's target interest rate.

Most countries have fractional reserve banking systems in which each bank must hold reserves (vault cash plus deposits at the central bank) at least equal to the required reserve ratio times its customer deposits. Banks with excess reserves can lend them to banks that need

reserves to meet their reserve requirements. The central bank can increase the money supply by (1) buying securities from banks, (2) lowering the required reserve ratio, and/or (3) reducing its target for the interest rate at which banks borrow and lend reserves among themselves. In each case, the opposite action would decrease the money supply.

When the central bank buys securities from banks in an open-market operation, it pays for them with a corresponding increase in bank reserves. This increases the amount of deposits banks can accept from their customers—that is, the money supply. Similarly, cutting the required reserve ratio increases the level of deposits (i.e., money) consistent with a given level of reserves in the system. If the central bank chooses to target an interbank lending rate, as the Federal Reserve targets the federal funds rate in the United States, then it must add or drain reserves via open-market operations to maintain the target interest rate. If it raises its target interest rate, it will have to drain reserves in order to make reserves more expensive in the interbank market; if it lowers the rate, it will add reserves, which will become less expensive. Thus, open-market operations and interest rate targeting are very closely related. The main distinction is whether the central bank chooses to target a level of reserves and let the market determine the interest rate or it chooses to target the interest rate and let the market (banks) determine the level of reserves they desire to hold at that rate.

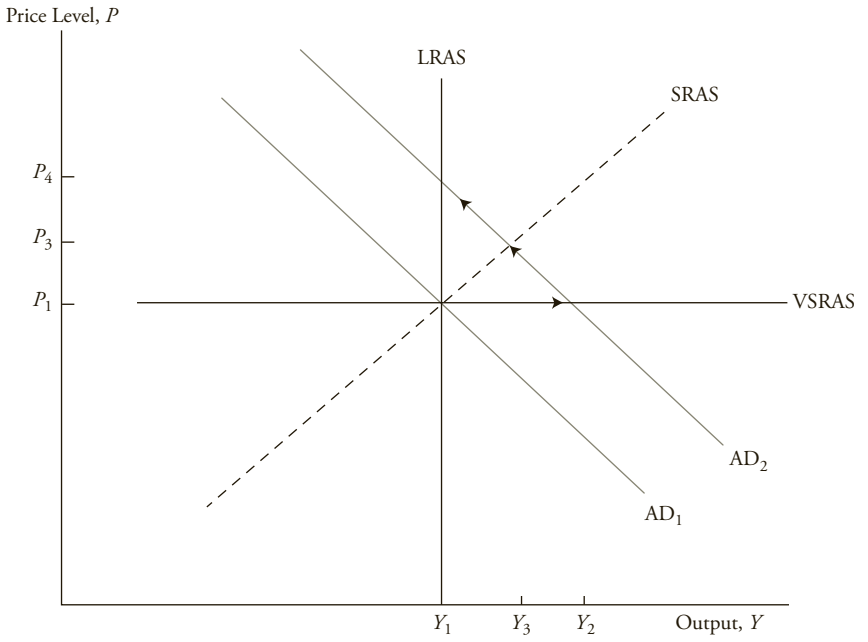
An increase in the money supply shifts the AD curve to the right so that each price level corresponds to a higher level of income and expenditure.[12] There are various channels through which the additional expenditures may be induced. For example, the interest rate reduction required to induce investors to hold the additional money balances will encourage companies to invest more and households to borrow to purchase durable goods, such as cars. In addition, banks may facilitate greater expenditure by raising credit limits and loosening credit standards. Conversely, a reduction in the money supply shifts the AD curve to the left.

Exhibit 5-17 illustrates the short-run and long-run effect of an expansionary monetary policy. Suppose the central bank expands the money supply in an attempt to stimulate demand when the economy is already in long-run equilibrium. The expansionary policy will shift the AD curve to the right, from AD_1 to AD_2. In the very short run, output will expand from Y_1 to Y_2 without an increase in the price level. After operating at higher-than-normal production rates for a few months or quarters, companies will begin to push for price increases, and input prices will begin to rise as well. The aggregate supply curve will steepen, and prices will increase to P_3 while output declines to Y_3. As input prices become more flexible, the AS curve will steepen until, in the long run, it is vertical and output has returned to the long-run natural level, Y_1, with prices rising to P_4. Thus, expanding the money supply increases output in the short run, but in the long run it affects only the price level.

3.3.1.6. Exchange Rate An exchange rate is the price of one currency relative to another. Changes in the exchange rate affect the price of exports and imports and thus aggregate demand. For example, a lower euro relative to other currencies makes European exports cheaper in world markets and foreign products sold in Europe (European imports) more expensive. Therefore, a lower euro should cause European exports to increase and imports to decline, causing the AD curve to shift to the right. Conversely, a stronger euro reduces exports and raises imports, and the AD curve shifts to the left.

[12]An unusual but important special case known as a liquidity trap occurs if (1) banks are willing to hold virtually unlimited excess reserves rather than expand their balance sheets by taking deposits and making loans and/or (2) demand for money balances by households and companies is insensitive to the level of income. In a liquidity trap, monetary policy will be ineffective and the AD curve will not shift despite the central bank's efforts. Some have argued that this was a reasonable description of the U.S. situation in 2010.

EXHIBIT 5-17 Short-Run and Long-Run Effect of Monetary Expansion



3.3.1.7. Growth in the Global Economy International trade is what links countries together and creates a global economy. Faster economic growth in foreign markets encourages foreigners to buy more products from domestic producers and increases exports. For example, rapid GDP growth in China has increased Chinese demand for foreign products. Japan has benefited from this rapid growth because it has exported more products to China. In terms of the AD and AS model, the AD curve for Japan has shifted to the right because of increased demand for Japanese products in China, resulting in higher exports. A decline in the growth rate of China's economy would have a negative effect on the Japanese economy because exports would be lower. This would cause the Japanese AD curve to shift to the left.

What happens to interest rates when the AD curve shifts? In the case of an increase in the money supply, the interest rate declines at each price level because the increase in income (Y) increases saving and rates must decline to induce a corresponding increase in investment spending (I). In each of the other cases considered earlier, a rightward shift in the AD curve will increase the interest rate at each price level. With the real money supply held constant, the interest rate must rise as income increases. The increase in the interest rate reduces the demand for money at each level of expenditure/income and, therefore, allows expenditure/income to increase without an increase in the money supply. In terms of the quantity theory of money equation, this corresponds to a higher velocity of money, V.

The main factors that shift the AD curve are summarized in Exhibit 5-18. In each case, the impact of the factor is considered in isolation. In practice, however, various factors may be at work simultaneously and may interact. This is especially true with regard to expectational factors—consumer and business confidence—which are likely to be influenced by other developments.

EXHIBIT 5-18 Impact of Factors Shifting Aggregate Demand

An Increase in These Factors	Shifts the AD Curve	Reason
Stock prices	Rightward: Increase in AD	Higher consumption
Housing prices	Rightward: Increase in AD	Higher consumption
Consumer confidence	Rightward: Increase in AD	Higher consumption
Business confidence	Rightward: Increase in AD	Higher investment
Capacity utilization	Rightward: Increase in AD	Higher investment
Government spending	Rightward: Increase in AD	Government spending a component of AD
Taxes	Leftward: Decrease in AD	Lower consumption and investment
Bank reserves	Rightward: Increase in AD	Lower interest rates, higher investment, and possibly higher consumption
Exchange rate (foreign currency per unit of domestic currency)	Leftward: Decrease in AD	Lower exports and higher imports
Global growth	Rightward: Increase in AD	Higher exports

EXAMPLE 5-8 Shifts in Aggregate Demand

François Ubert is a portfolio manager with EuroWorld, a French investment management firm. Ubert is considering increasing his clients' portfolio exposure to Brazilian equities. Before doing so, he asks you to prepare a report on the following recent economic events in Brazil and to summarize the impact of each event on the Brazilian economy and on Brazilian equity and fixed-income securities.

1. The Brazilian central bank reduced bank reserves, resulting in a lower money supply.
2. The capacity utilization rate in Brazil is currently estimated to be 86.4 percent, a 2.7 percent increase from the previous year.
3. Corporate profits reported by Brazilian companies increased by 30 percent over last year's levels, and corporations have revised their forecasts of future profitability upward.
4. The government recently announced that it plans to start construction on a number of hydroelectric projects to reduce Brazil's reliance on imported oil.
5. Forecasts by private-sector economists project that the European economy will enter a recession in the next year.

Solution to 1: This monetary policy action is designed to reduce consumption and business investment spending. The reduction in real money balances will increase

interest rates and discourage lending within the banking system. Higher interest rates and tighter credit will reduce both investment and consumption expenditures and shift the AD curve to the left. The prices of fixed-income securities will fall because of the rise in interest rates. The reduction in aggregate output should lower corporate profits, and it is likely that equity prices will also fall.

Solution to 2: Capacity utilization is a key factor determining the level of investment spending. A current utilization rate of over 86 percent and an increase from the previous year indicate a growing lack of spare capacity in the Brazilian economy. As a result, businesses will probably increase their level of capital spending. This will increase AD and shift the AD curve to the right. Higher economic activity (income/output) will cause upward pressure on interest rates and may have a negative impact on fixed-income securities. Higher income/output should increase corporate profits and is likely to have a positive impact on equity securities.

Solution to 3: Expected corporate profits are an important determinant of the level of investment spending. The large increase in expected profits will raise the level of investment spending and increase aggregate demand. This will shift the AD curve to the right. The increase in corporate profits and the resulting increase in economic output should have a positive impact on equities. The increase in output will put upward pressure on interest rates and downward pressure on the prices of fixed-income securities.

Solution to 4: Fiscal policy uses government spending to influence the level and growth rate of economic activity. The announcement indicates an increase in government spending, which is a direct component of AD. Therefore, higher spending on the projects will increase AD and shift the AD curve to the right. The increase in output and expenditure should be positive for equities. But it will be negative for existing fixed-income investments because higher interest rates will be required to induce investors to buy and hold the government debt issued to fund the new projects.

Solution to 5: A recession in Europe will decrease the demand for Brazilian exports by European households and businesses and shift the AD curve to the left. The resulting decline in income and downward pressure on prices will be positive for fixed-income securities but negative for equities.

3.3.2. Shifts in Short-Run Aggregate Supply

Factors that change the cost of production or expected profit margins will cause the SRAS curve to shift. These factors include changes in:

- Nominal wages.
- Input prices, including the price of natural resources.
- Expectations about future output prices and the overall price level.
- Business taxes and subsidies.
- The exchange rate.

In addition, factors that shift the long-run AS curve (see Section 3.3.3) will also shift the SRAS curve by a corresponding amount because the SRAS and LRAS reflect the same underlying resources and technology. As the economy's resources and technology change, the full employment (or natural) level of output changes, and both the LRAS and SRAS shift accordingly.

3.3.2.1. Change in Nominal Wages Changes in nominal wages shift the short-run AS curve because wages are often the largest component of a company's costs. An increase in nominal wages raises production costs, resulting in a decrease in AS and a leftward shift in the SRAS curve. Lower wages shift the AS curve to the right. It is important to note that changes in nominal wages have no impact on the LRAS curve.

A better way to measure the impact of labor costs on the AS curve is to measure the change in unit labor cost. We define the change in unit labor cost as:

$$\begin{aligned}\%\text{ Change in unit labor cost} &= \%\text{ Change in nominal wages} \\ &\quad - \%\text{ Change in productivity}\end{aligned}$$

EXAMPLE 5-9 Unit Labor Cost and Short-Run Aggregate Supply

Suppose Finnish workers are paid €20 an hour and are able to produce 100 cell phones in an hour. The labor cost per cell phone is €0.20 (€20 divided by 100 units). If the wages per hour for Finnish workers rise by 10 percent, from €20 to €22, and they are able to raise their productivity by 10 percent, what is the impact on unit labor cost and the short-run aggregate supply curve?

Solution: The workers can now produce 110 cell phones per hour, and unit labor cost will not change (22/110 = 0.20). In this case, the SRAS curve will remain in its original position. If wages had increased by 20 percent instead of 10 percent, then unit labor cost would have increased and the SRAS would shift to the left. Conversely, if the wage increase were only 5 percent, then unit labor cost would have decreased and the SRAS would shift to the right.

3.3.2.2. Change in Input Prices The price of raw materials is an important component of cost for many businesses. Lower input prices reduce the cost of production, which, in turn, makes companies willing to produce more at any output price. This is reflected in a rightward shift of the SRAS curve. Conversely, higher input prices increase production costs, which, in turn, cause companies to reduce production at any output price. This shifts the SRAS curve to the left. During the 1970s, high oil prices caused the SRAS curve in most countries to shift to the left. In contrast, in the mid-1980s, declining oil prices lowered the cost of production and shifted the SRAS curve in most countries to the right. Oil prices currently have a smaller impact on the global economy than in the 1970s and 1980s because most countries have reduced their reliance on oil and improved their energy efficiency so that they now use less energy per unit of GDP.

3.3.2.3. Change in Expectations about Future Prices The impact of expected future prices on current output decisions is not as straightforward as it might seem. First, each company is primarily concerned about the price of its own output rather than the general price level. The latter may be more reflective of its costs. If it expects its own output price to rise relative to the general price level, then it may increase production in response to the perceived change in its profit margin; if it expects its output price to fall, it may decrease production. As more and more companies become optimistic about their ability to raise the relative price of their product, the SRAS will shift to the right; if pessimistic, the SRAS will shift to the left. In the aggregate, of course, companies can neither raise nor lower their prices relative to the general price level. Hence, shifts in the SRAS driven by such price expectations are likely to be modest and temporary. Second, considering future prices introduces a temporal aspect into decision making. If the future price level is expected to be higher, companies may decide to produce more today in order to expand inventory available for future sale. But they will do so only if the cost of carrying inventory (financing, storage, and spoilage) is less than they expect to save on production costs by producing more today and less in the future. Conversely, they may cut current production and sell out of existing inventory if they expect future prices (and costs) to be lower.

The upshot is that expectations of higher future prices are likely to shift the SRAS curve to the right, and lower expectations probably will shift the curve to the left; but either impact may be modest and/or temporary.

3.3.2.4. Change in Business Taxes and Subsidies Higher business taxes increase production costs per unit and shift the short-run AS curve to the left. Business subsidies are a payment from the government to the producer. Subsidies for businesses lower their production costs and shift the SRAS curve to the right.

3.3.2.5. Change in the Exchange Rate Many countries import raw materials, including energy and intermediate goods. As a result, changes in the exchange rate can affect the cost of production and, therefore, aggregate supply. A higher yen relative to the euro will lower the cost of raw materials and intermediate goods imported to Japan from Europe. This, in turn, will lower the production costs of Japanese producers and shift the AS curve in Japan to the right. A lower yen will have the opposite effect.

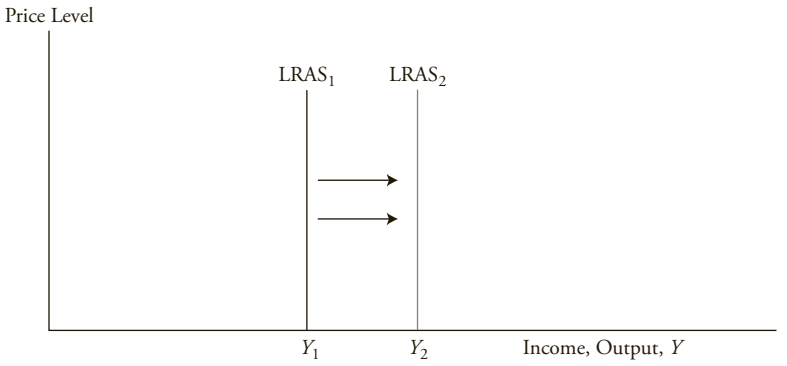
3.3.3. Shifts in Long-Run Aggregate Supply

As discussed earlier, the position of the LRAS curve is determined by the potential output of the economy. **Potential GDP** measures the productive capacity of the economy and is the level of real GDP that can be produced at full employment. Potential GDP is not a static concept but can increase each year at a steady rate as the economy's resource capacity grows. Therefore, any factor increasing the resource base of an economy causes the LRAS curve to shift as shown in Exhibit 5-19.

These factors include changes in:

- Supply of labor and quality of labor forces (human capital).
- Supply of natural resources.
- Supply of physical capital.
- Productivity and technology.

EXHIBIT 5-19 Shift in Long-Run Aggregate Supply (LRAS) Curve



3.3.3.1. Supply of Labor The larger the supply of labor, the more output the economy can produce. The labor supply depends on growth in the population, the labor force participation rate (the percentage of the population working or looking for work), and net immigration. The determinants of the labor supply are discussed in more detail in Section 4. Increases in the labor supply shift the LRAS curve to the right. Decreases shift the curve to the left.

3.3.3.2. Supply of Natural Resources Natural resources are essential inputs to the production process and include everything from available land to oil to water. Increased availability of natural resources shifts the LRAS curve to the right.

3.3.3.3. Supply of Physical Capital Investment in new property, plant, equipment, and software is an essential ingredient for growth. An increase in the stock of physical capital will increase the capacity of the economy to produce goods and services. Simply put, if workers are provided with more and better equipment to use, they should be able to produce more output than they could with the older equipment. Thus, strong growth in business investment, which increases the supply of physical capital, shifts the LRAS curve to the right.

3.3.3.4. Supply of Human Capital Another way to raise the productive capacity of a country is to increase human capital—the quality of the labor force—through training, skills development, and education. Improvement in the quality of the labor force shifts the LRAS curve to the right.

3.3.3.5. Labor Productivity and Technology Another important factor affecting the productive capacity of an economy is how efficient labor is in transforming inputs into final goods and services. **Productivity** measures the efficiency of labor and is the amount of output produced by workers in a given period of time—for example, output per hour worked. An increase in productivity decreases labor cost, improves profitability, and results in higher output. Two of the main drivers of labor productivity—physical capital per worker and the quality of the workforce—have been discussed. The third key determinant of productivity is technology. Advances in technology shift the LRAS curve to the right.

EXAMPLE 5-10 Unit Labor Cost and Long-Run Aggregate Supply

Again, suppose Finnish workers are paid €20 per hour and are able to produce 100 cell phones in an hour. If workers develop a new technique for assembly and are able to produce 200 cell phones per hour, what is the impact on the long-run aggregate supply curve?

Solution: Labor cost per unit will decline to €0.10 (20/200 = €0.10 per cell phone). As a result, profit per unit will rise and companies will have an incentive to increase production. Thus, the LRAS curve shifts to the right.

The factors shifting the AS curve are summarized in Exhibit 5-20. Rightward shifts in the SRAS or LRAS curves are defined as an increase in supply. Leftward shifts in the SRAS or LRAS curves represent a decrease in supply.

As with our summary of factors that shift the AD curve, Exhibit 5-20 considers each of the factors affecting aggregate supply in isolation. In practice, various factors will be at work simultaneously and may interact. This is especially important with respect to interaction between factors listed as affecting only SRAS and those that also impact LRAS.

For example, consider an increase in the cost of inputs (e.g., energy). This shifts the SRAS curve to the left, but according to Exhibit 5-20, it has no effect on LRAS. This presumes that

EXHIBIT 5-20 Impact of Factors Shifting Aggregate Supply

An Increase in These Factors	Shifts SRAS	Shifts LRAS	Reason
Supply of labor	Rightward	Rightward	Increases resource base
Supply of natural resources	Rightward	Rightward	Increases resource base
Supply of human capital	Rightward	Rightward	Increases resource base
Supply of physical capital	Rightward	Rightward	Increases resource base
Productivity and technology	Rightward	Rightward	Improves efficiency of inputs
Nominal wages	Leftward	No impact	Increases labor cost
Input prices (e.g., energy)	Leftward	No impact	Increases cost of production
Expectation of future prices	Rightward	No impact	Anticipation of higher costs and/or perception of improved pricing power
Business taxes	Leftward	No impact	Increases cost of production
Subsidy	Rightward	No impact	Lowers cost of production
Exchange rate	Rightward	No impact	Lowers cost of production

there has not been a permanent change in the relative prices of the factors of production. If there has been a permanent change, companies will be forced to conserve on the now more expensive input and will not be able to produce as efficiently. The LRAS curve would, therefore, shift to the left, just as it would if the available supply of natural resources had declined relative to the supply of other inputs. Indeed, that is the most likely cause of a permanent change in relative input prices.

EXAMPLE 5-11 Shifts in Aggregate Supply

Jane Donovan is a portfolio manager for a global mutual fund. Currently, the fund has 10 percent of its assets invested in Chinese equities. She is considering increasing the fund's allocation to the Chinese equity market. The decision will be based on an analysis of the following economic developments and their impact on the Chinese economy and equity market. What is the impact on SRAS and LRAS from the following factors?

1. Global oil prices, currently near their longer-run trend at $75 a barrel, have increased from $35 a barrel over the past three years because of strong demand from China and India.
2. The number of students studying engineering has dramatically increased at Chinese universities over the past decade.
3. Wages for China's workers are rising, leading some multinational companies to consider shifting their investments to Vietnam or India.
4. Recent data show that business investment as a share of GDP is over 40 percent in China.
5. The Bank of China is likely to permit the yuan to appreciate by 10 percent over the next year.

Solution to 1: Higher energy prices cause a decrease in short-run AS and shift the SRAS curve to the left. Because oil prices are back to their longer-run trend, the leftward shift in SRAS essentially reverses a previous shift that occurred when oil prices fell to $35, and it is likely that there will be no impact on the LRAS curve. Lower output and profit are likely to have a negative effect on Chinese equity prices.

Solution to 2: More students studying engineering indicates an improvement in the quality of the labor force—an increase in human capital. As a result, AS increases and the AS curve shifts to the right. Both short-run and long-run curves are affected. Higher output and profits may be expected to have a positive effect on Chinese equity prices.

Solution to 3: The increase in wages increases labor costs for businesses, causes short-run aggregate supply to decline, and shifts the SRAS curve to the left. Lower output and profit should have a negative effect on Chinese equity prices.

Solution to 4: The high level of business investment indicates that the capital stock in China is growing at a fast rate. This means that workers have more capital to use, which

increases their productivity. Thus, AS increases and the AS curve shifts to the right. Both short-run AS and long-run AS are affected. Higher output should have a positive effect on Chinese equity prices.

Solution to 5: The probable appreciation of the yuan means that the cost of imported raw materials, such as iron ore, coal, and oil, will be lower for Chinese companies. As a result, short-run AS increases and the SRAS curve shifts to the right. The LRAS curve may also shift to the right if the appreciation of the yuan is permanent and global commodity prices do not fully adjust. Higher output and profit should have a positive effect on Chinese equity prices.*

The implications of these five factors for equity investment in China are ambiguous. If the long-run effects dominate, however, then the net impact should be positive. The positive factors—the high level of investment and the growing pool of engineering students—have a lasting impact on output and profit. The negative factors—higher wages and oil prices—should be temporary because wages will realign with the price level and the increase in oil prices appears to offset a previous temporary decline. The reduction in raw material prices due to the stronger currency is positive for output, profit, and equities in the short run and perhaps in the long run as well.

*The alert reader may have noted that the stronger yuan will also reduce export demand and shift the AD curve to the left. The combined impact of the AD and AS shifts on output, profit, and equity prices is ambiguous.

3.4. Equilibrium GDP and Prices

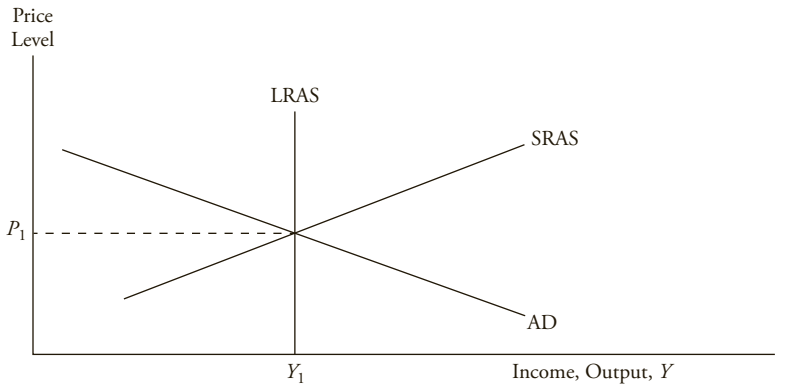
Now that we have discussed the components of the AD and AS model, we can combine them to determine the real level of GDP and the price level. Equilibrium occurs where the AD and AS curves intersect. At this point, the quantity of aggregate output demanded (or the level of aggregate expenditures) is equal to the quantity of aggregate output supplied. In Exhibit 5-21, equilibrium price and GDP occur at P_1 and Y_1. If the price level is above P_1, then the quantity of output supplied exceeds the amount demanded. This situation would result in unsold inventories and would require a reduction in production and in prices. If the price level is below P_1, then the quantity of aggregate output demanded exceeds the quantity of aggregate output supplied. This situation would result in a shortage of goods that would put upward pressure on prices.

It is important to understand that short-run macroeconomic equilibrium may occur at a level above or below full employment. We consider four possible types of macroeconomic equilibrium:

1. Long-run full employment
2. Short-run recessionary gap
3. Short-run inflationary gap
4. Short-run stagflation

From an investment perspective, the performance of asset classes and financial markets will differ in each of the above cases as the economy makes the adjustment toward the macroeconomic equilibrium. We look at these differences later in the chapter.

EXHIBIT 5-21 Long-Run Macroeconomic Equilibrium



3.4.1. Long-Run Equilibrium

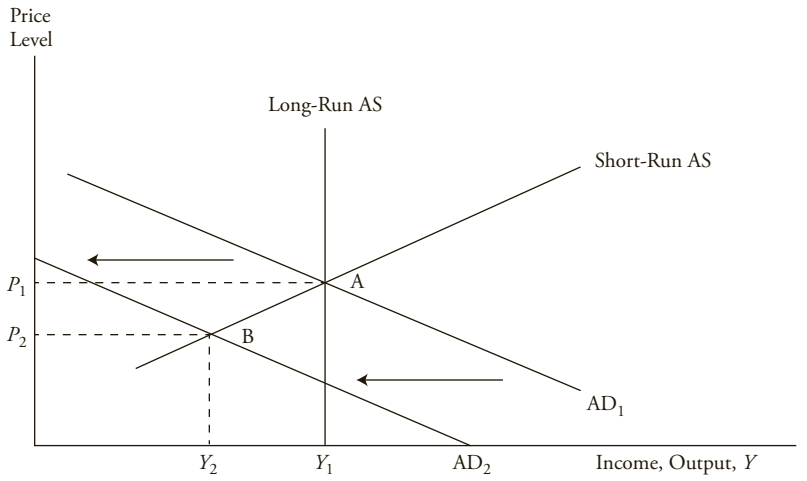
Exhibit 5-21 shows the long-run full employment equilibrium for an economy. In this case, equilibrium occurs where the AD curve intersects the SRAS curve at a point on the LRAS curve. Because equilibrium occurs at a point on the LRAS curve, the economy is at potential real GDP. Both labor and capital are fully employed, and everyone who wants a job has one. *In the long run, equilibrium GDP is equal to potential GDP.*

In practice, the level of potential GDP is difficult to measure with precision. Because of fluctuations arising from shifts in the AD and SRAS curves, the economy rarely operates at potential GDP. Thus, potential GDP is not observable from the data on actual GDP. In addition, potential GDP is determined by factors that are themselves difficult to measure (see Section 4.2). Thus, bottom-up estimates of the *level* of potential output are also quite imprecise. However, as will be discussed in Section 4, economists have confidence that the long-run *growth rate* of potential GDP can be estimated well enough to provide meaningful guidance for analysts and policy makers. Hence, in the short run, economists generally focus on factors that cause actual GDP to grow faster or slower than their estimate of the long-run growth rate of potential output. In addition, they focus on measures that indicate, albeit imprecisely, the extent to which the economy is operating above or below its productive capacity, such as unemployment and capacity utilization.

3.4.2. Recessionary Gap

Cyclical fluctuations in real GDP and prices are caused by shifts in both the AD curve and the SRAS curve. A decline in AD or a leftward shift in the AD curve results in lower GDP and lower prices. Such declines in AD lead to economic contractions, and if such declines drive demand below the economy's potential GDP, the economy goes into a recession. In Exhibit 5-22, when aggregate demand falls, the equilibrium shifts from point A to point B. Real GDP contracts from Y_1 to Y_2, and the aggregate price level falls from P_1 to P_2. Because of the decline in demand, companies reduce their workforces and the unemployment rate rises.

EXHIBIT 5-22 Recessionary Gap



The economy is in **recession**,[13] and the recessionary gap is measured as the difference between Y_2 and Y_1 or the amount by which equilibrium output is below potential GDP. Thus, a recessionary gap occurs when the AD curve intersects the short-run AS curve at a short-run equilibrium level of GDP below potential GDP. *Most importantly, in contrast to full employment, equilibrium GDP is below potential GDP.*

Any of the factors discussed in Section 3.3.1 could cause the shift in the AD curve. Tightening of monetary policy, higher taxes, more pessimistic consumers and businesses, and lower equity and housing prices all reduce AD and are all possible causes of a recession.

The question is: How does the economy return to full employment? There is considerable debate among economists about the answer to this question. Some economists argue that an automatic, self-correcting mechanism will push the economy back to its potential, without the need for government action. The idea is that because of the decline in prices and higher unemployment, workers will be willing to accept lower nominal wages. Workers will do this because each currency unit of wages now buys more goods and services because of their lower prices. As a result, lower wages and input prices will cause the SRAS curve to shift to the right (see Exhibit 5-20) and push the economy back to full employment and potential GDP.

The problem is that this price mechanism can take several years to work. As an alternative, government can use the tools of fiscal and monetary policy to shift the AD curve to the right (from point B to point A in Exhibit 5-22) and move the economy back to full employment. On the fiscal side, policy makers can reduce taxes or increase government spending. On the monetary side, the central bank can lower interest rates or increase the money supply. The problem, however, is that variable lags in the effectiveness of these policy measures imply that policy adjustments may end up reinforcing rather than counteracting underlying shifts in the economy.

[13]A recession is defined as a period during which real GDP decreases (i.e., negative growth) for at least two successive quarters or a period of significant decline in total output, income, employment, and sales usually lasting from six months to a year.

3.4.2.1. Investment Implications of a Decrease in AD Aggregate demand and aggregate supply are theoretical measures that are very hard to measure directly. Most governments, however, publish statistics that provide an indication of the direction in which aggregate demand and supply are moving over time. For example, statistics on consumer sentiment, factory orders for durable and nondurable goods, the value of unfilled orders, the number of new housing starts, the number of hours worked, and changes in inventories provide an indication of the direction of aggregate demand. If these statistics suggest that a recession is caused by a decline in AD, the following conditions are likely to occur:

- Corporate profits will decline.
- Commodity prices will decline.
- Interest rates will decline.
- Demand for credit will decline.

This suggests the following investment strategy:

- Reduce investments in **cyclical companies**[14] because their earnings are likely to decline the most in an economic slowdown.
- Reduce investments in commodities and commodity-oriented companies because the decline in commodity prices will slow revenue growth and reduce profit margins.
- Increase investments in **defensive companies**[15] because they are likely to experience only modest earnings declines in an economic slowdown.
- Increase investments in investment-grade or government-issued fixed-income securities. The prices of these securities should increase as interest rates decline.
- Increase investments in long-maturity fixed-income securities because their prices will be more responsive to the decline in interest rates than the prices of shorter-maturity securities will be.
- Reduce investments in speculative equity securities and in fixed-income securities with low credit quality ratings.

As with most investment strategies, this strategy will be most successful if it is implemented before other market participants recognize the opportunities and asset prices adjust.

EXAMPLE 5-12 Using AD and AS: The Recession of 2007–2009

Many Asian economies were more adversely affected than the United States by the global recession that began in late 2007. In the first quarter of 2009, real GDP fell at an annualized rate of 16 percent in Japan, 11 percent in Singapore, and 9 percent in

[14]Cyclical companies are companies with sales and profits that regularly expand and contract with the business cycle or state of the economy (for example, automobile and chemical companies).
[15]Defensive companies are companies with sales and profits that have little sensitivity to the business cycle or state of the economy (for example, food and pharmaceutical companies).

Taiwan, compared with a 6 percent annualized decline in the United States. Using the data on exports as a share of GDP shown in Exhibit 5-23, explain how the following economic factors contributed to the recession in the Asian economies:

1. Collapse of house prices and home construction in the United States.
2. Oil prices rising from around $30 a barrel in 2004 to nearly $150 a barrel in 2008. (*Note:* Most of the markets in eastern Asia, such as Hong Kong, Japan, South Korea, and Taiwan, rely on imports for almost all of their oil and energy needs. In contrast, the United States has a large domestic energy industry and imports about one-half of its oil.)
3. The dramatic reduction in credit availability following the collapse or near collapse of major financial institutions in 2008.

EXHIBIT 5-23 Exports as a Share of GDP, 2007

Market	Exports as a Percentage of GDP	Percentage of Exports Going to United States
Singapore	186%	11.2%
Hong Kong	166	11.5
Taiwan	62	11.6
South Korea	53	10.9
Germany	47	7.1
China	37	26.4
Mexico	28	80.2
Kenya	27	8.2
Japan	17	20.1
India	14	17.0
United States	12	—
Ethiopia	11	6.7

Sources: World Bank: World Development Indicators; OECD.StatExtracts (http://stats.oecd.org).

Solution to 1: The collapse in housing prices caused housing construction spending, a component of business investment, to decline in the United States. The decline in housing prices also caused a sharp fall in household wealth. As a result, consumption spending in the United States declined because of the wealth effect. The decline in both consumption and housing construction shifted the AD curve for the United States to the left, resulting in a U.S. recession. The link to the Asian economies was through global trade because exports represent such a large share of the Asian countries' GDPs. In 2007, exports as a share of GDP (Exhibit 5-23) were 186 percent in Singapore, 166 percent in Hong Kong, 62 percent in Taiwan, 53 percent in South Korea, and

37 percent in China. In turn, each of these markets exports a significant amount of goods and services to the United States; for example, over 26 percent of Chinese exports were going to the United States. Thus, the recession in the United States and especially the decline in U.S. consumption spending caused a sharp fall in exports from these countries. This lowered their AD and caused the AD curve to shift to the left, resulting in a recessionary gap in these countries.

Solution to 2: The rise in oil prices increased input cost and shifted the short-run AS curve to the left. Because the eastern Asian economies are heavily dependent on imported oil, their economies were more adversely affected than was the economy of the United States.

Solution to 3: The decline in housing prices caused financial institutions in the United States to suffer large losses on housing-related loans and securities. Several large lenders collapsed, and the U.S. Treasury and the Federal Reserve had to intervene to prevent a wave of bankruptcies by large financial institutions. As a result of the crisis, it became difficult for households and businesses to obtain credit to finance their spending. This caused AD to fall and increased the severity of the recession in the United States, resulting in a significant decline in U.S. imports and thus exports from the Asian countries. In addition, the financial crisis made it more difficult to get trade finance, further reducing exports from Asia.

In summary, global investors need to be aware of the growing linkages among countries and the extent that one country's economic growth depends on demand from outside as well as from within that country. Data on exports as a percentage of a country's GDP provide an indication of this dependence. Although Japan is often viewed as an export-driven economy, Exhibit 5-23 shows that exports are only 17 percent of its GDP. Similarly, the economy of India depends largely on domestic spending for growth because exports account for only 14 percent of GDP.

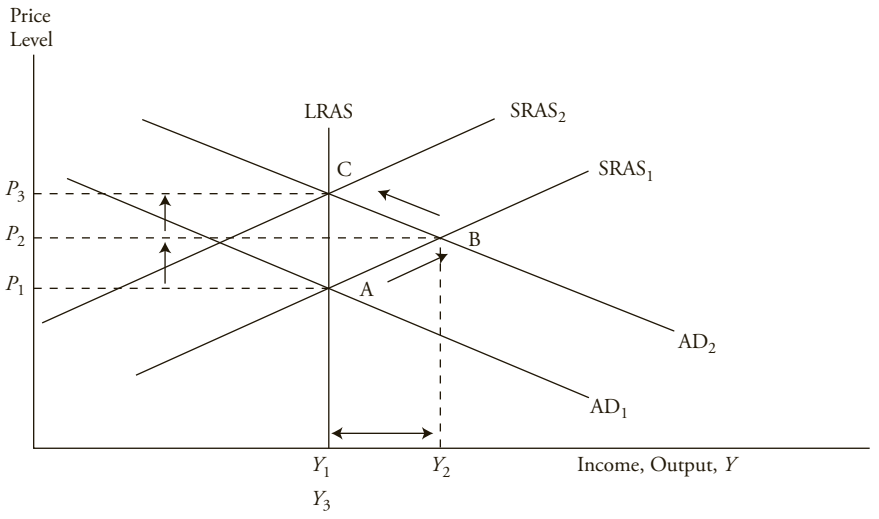
3.4.3. Inflationary Gap

Increases in AD lead to economic expansions as real GDP and employment increase. If the expansion drives the economy beyond its production capacity, however, **inflation**[16] will occur. As summarized in Exhibit 5-18, higher government spending, lower taxes, a more optimistic outlook among consumers and businesses, a weaker domestic currency, rising equity and housing prices, and an increase in the money supply would each stimulate aggregate demand and shift the AD curve to the right. If aggregate supply does not increase to match the increase in AD, a rise in the overall level of prices will result.

In Exhibit 5-24, an increase in AD will shift the equilibrium level of GDP from point A to point B. Real output increases from Y_1 to Y_2, and the aggregate price level rises from P_1 to P_2. As a result of the increase in aggregate demand, companies increase their production and hire more workers. The unemployment rate declines. Once an economy reaches its potential GDP, however, companies must pay higher wages and other input prices to further increase

[16]The inflation rate is defined as the increase in the general price level from one period to the next.

EXHIBIT 5-24 Inflationary Gap



production. The economy now faces an inflationary gap, measured by the difference between Y_2 and Y_1 in Exhibit 5-24. *An inflationary gap occurs when the economy's short-run level of equilibrium GDP is above potential GDP, resulting in upward pressure on prices.*

GDP cannot remain at Y_2 for long because the economy is overutilizing its resources; that is, extra shifts of workers are hired and plants and equipment are operating at their maximum capacity. Eventually, workers become tired and plants and equipment wear out. The increase in the general price level and input prices will set in motion the process of returning the economy back to potential GDP. Higher wages and input prices shift the SRAS supply curve to the left (from $SRAS_1$ to $SRAS_2$), moving the economy to point C in Exhibit 5-24. Again, this self-correcting mechanism may work slowly.

A nation's government and its central bank can attempt to use the tools of fiscal and monetary policy to control inflation by shifting the AD curve to the left (AD_2 to AD_1 in Exhibit 5-24) so that the return to full employment occurs without the price increase. From a fiscal perspective, policy makers can raise taxes or cut government spending. From a monetary perspective, the central bank can reduce bank reserves, resulting in a decrease in the growth of the money supply and higher interest rates.

3.4.3.1. Investment Implications of an Increase in AD Resulting in an Inflationary Gap If economic statistics (consumer sentiment, factory orders for durable and nondurable goods, etc.) suggest that there is an expansion caused by an increase in AD, the following conditions are likely to occur:

- Corporate profits will rise.
- Commodity prices will increase.
- Interest rates will rise.
- Inflationary pressures will build.

This suggests the following investment strategy:

- Increase investment in cyclical companies because they are expected to have the largest increase in earnings.
- Reduce investments in defensive companies because they are expected to have only a modest increase in earnings.
- Increase investments in commodities and commodity-oriented equities because they will benefit from higher production and output.
- Reduce investments in fixed-income securities, especially longer-maturity securities, because they will decline in price as interest rates rise. Raise exposure to speculative fixed-income securities (junk bonds) because default risks decrease in an economic expansion.

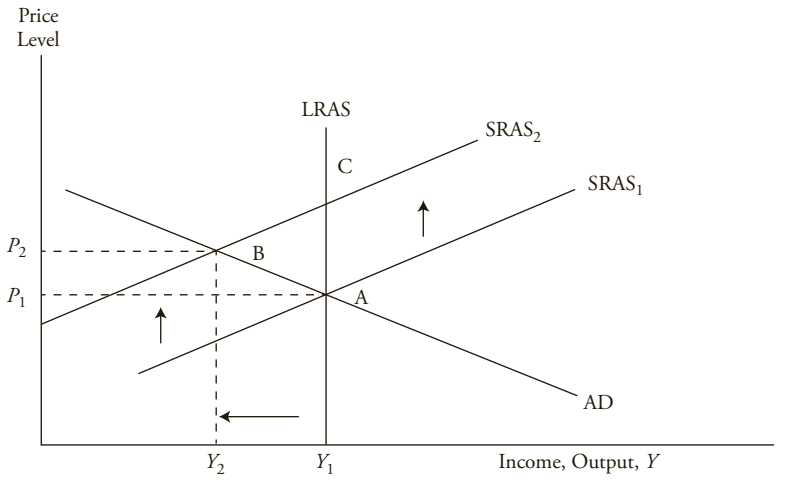
3.4.4. Stagflation: Both High Inflation and High Unemployment

Structural fluctuations in real GDP are caused by fluctuations in SRAS. Declines in aggregate supply bring about **stagflation**—high unemployment and increased inflation. Increases in aggregate supply conversely give rise to high economic growth and low inflation.

Exhibit 5-25 shows the case of a decline in aggregate supply, perhaps caused by an unexpected increase in basic material and oil prices. The equilibrium level of GDP shifts from point A to point B. The economy experiences a recession as GDP falls from Y_1 to Y_2, but the price level, instead of falling, rises from P_1 to P_2. Over time, the reduction in output and employment should put downward pressure on wages and input prices and shift the SRAS curve back to the right, reestablishing full employment equilibrium at point A. However, this mechanism may be painfully slow. Policy makers may use fiscal and monetary policy to shift the AD curve to the right, as previously discussed, but at the cost of a permanently higher price level at point C.

The global economy experienced stagflation in the mid-1970s and early 1980s. Both unemployment and inflation soared. The problem was caused by a sharp decline in aggregate supply fueled by higher input prices, especially the price of oil. In 1973, the price of oil quadrupled. A steep global recession began in late 1973 and lasted through early 1975. The

EXHIBIT 5-25 Stagflation



recession was unusual because prices rose rather than declined as would be expected in a typical demand-caused downturn. In 1979–1980, the price of oil doubled. Higher energy prices shifted the SRAS curve to the left, as shown in Exhibit 5-25, leading to a global recession from 1980 to 1982. In the United States, the contraction in output was reinforced by the Federal Reserve's decision to tighten monetary policy to fight the supply-induced inflation.

3.4.4.1. Investment Implications of Shift in AS Labor and raw material costs, including energy prices, determine the direction of shifts in short-run aggregate supply: Higher costs for labor, raw materials, and energy lead to a decrease in aggregate supply, resulting in lower economic growth and higher prices. Conversely, lower labor costs, raw material prices, and energy prices lead to an increase in aggregate supply, resulting in higher economic growth and a lower aggregate price level. Productivity is also an important factor. Higher rates of productivity growth shift the AS to the right, resulting in higher output and lower unit input prices. Lower rates of productivity growth do the opposite and shift the AS curve to the left.

From an investment perspective, a decline in AS (leftward shift of the SRAS curve) suggests:

- Reducing investment in fixed-income securities because rising output prices (i.e., inflation) put upward pressure on nominal interest rates.
- Reducing investment in most equity securities because profit margins are squeezed and output declines.
- Increasing investment in commodities or commodity-based companies because prices and profits are likely to rise.

In contrast, an increase in AS (rightward shift of the SRAS curve) due to higher productivity growth or lower labor, raw material, and energy costs is favorable for most asset classes other than commodities.

3.4.5. Conclusions on AD and AS

The business cycle and the resulting fluctuations in real GDP are caused by shifts in the economy represented by the AD and AS curves. The impact of these shifts can be summarized as follows:

- An increase in AD raises real GDP, lowers the unemployment rate, and increases the aggregate level of prices.
- A decrease in AD lowers real GDP, increases the unemployment rate, and decreases the aggregate level of prices.
- An increase in AS raises real GDP, lowers the unemployment rate, and lowers the aggregate level of prices.
- A decrease in AS lowers real GDP, raises the unemployment rate, and raises the aggregate level of prices.

If both curves shift, the effect is a combination of these individual effects. We can look at four possible cases:

1. *Both AD and AS increase.* If both AD and AS increase, real GDP will increase, but the impact on inflation is not clear unless we know the magnitude of the changes, because an increase in AD will increase the price level, whereas an increase in AS will decrease the price level. If AD increases more than AS, the price level will increase. If AS increases more than AD, however, the price level will decline.

2. *Both AD and AS decrease.* If both AD and AS decrease, real GDP and employment will decline, but the impact on the price level is not clear unless we know the magnitude of the changes, because a decrease in AD lowers the price level, whereas a decrease in AS raises the price level. If AD decreases more than AS, the price level will fall. If AS decreases more than AD, the price level will rise.
3. *AD increases and AS decreases.* If AD increases and AS declines, the price level will rise, but the effect on real GDP is not clear unless we know the magnitude of the changes, because an increase in AD increases real GDP, whereas a decrease in AS decreases real GDP. If AD increases more than AS declines, GDP will rise. If AS decreases more than AD increases, real GDP will fall.
4. *AD decreases and AS increases.* If AD decreases and AS increases, the price level will decline, but the impact on real GDP is not clear unless we know the magnitude of the changes, because a decrease in AD decreases real GDP, whereas an increase in AS increases real GDP. If AD decreases more than AS increases, real GDP will fall. If AS increases more than AD declines, real GDP will rise.

Exhibit 5-26 summarizes these four cases.

Whether the growth of the economy is demand driven or supply driven has an impact on asset prices. Demand-driven expansions are normally associated with rising interest rates and inflation, whereas contractions are associated with lower inflation and interest rates. Supply-driven expansions are associated with lower inflation and interest rates, whereas supply-driven contractions are associated with rising inflation and interest rates.

EXHIBIT 5-26 Effect of Combined Changes in AD and AS

Change in AD	Change in AS	Effect on Real GDP	Effect on Aggregate Price Level
Increase	Increase	Increase	Indeterminate
Decrease	Decrease	Decrease	Indeterminate
Increase	Decrease	Indeterminate	Increase
Decrease	Increase	Indeterminate	Decrease

EXAMPLE 5-13 Investment Strategy Based on AD and AS Curves

An analyst is evaluating the possibility of investing in China, Italy, Mexico, or Brazil. What are the equity and fixed-income investment opportunities in these countries based on the following events?

1. The Chinese government announced a spending plan of $1.2 trillion or 13 percent of GDP. In addition, the central bank of China eased monetary policy, resulting in a surge of lending.

2. The Italian government announced a decline in labor productivity, and it expects this trend to continue into the future.
3. In response to rising inflationary pressure, the Mexican central bank tightened monetary policy, and the government announced tax increases and spending cuts to balance the budget.
4. A major discovery of oil off the coast of Brazil lowered oil prices, while the Brazilian government announced a major increase in spending on public infrastructure to stimulate the economy.

Solution to 1: Stimulative fiscal and monetary policies should result in a demand-driven expansion. Investors should reduce investments in fixed-income securities and defensive companies and invest in cyclical companies and commodities. As a result, the prospects for growth-oriented equity investments look favorable in China.

Solution to 2: A decline in labor productivity will result in a decline in AS; that is, the AS curve will shift to the left. This is typically a poor investment environment. Investors should reduce investments in both fixed-income and equity securities and invest in commodities. Entry into Italian stocks and bonds does not look attractive.

Solution to 3: The policy measures put in place by the Mexican government and central bank will cause a drop in AD and likely result in a recession. Investors should increase their investments in fixed-income securities because interest rates will most likely decline as the recession deepens. This is a poor environment for equity securities.

Solution to 4: This is a situation where both the AD and AS curves will shift. The increase in spending on public infrastructure will shift the AD curve to the right, resulting in higher aggregate expenditures and prices. Lower oil prices will shift the AS curve to the right as well, resulting in higher GDP but lower prices. Thus, GDP will clearly increase, but the impact on prices and inflation is indeterminate. As a result, investors should increase their investment in equity securities; however, the impact on fixed-income securities is unclear.

Using AD and AS to Explain Japan's Economic Problem

Japan has experienced sluggish growth in real GDP for nearly two decades following the bursting of an asset and investment bubble in the late 1980s. At the same time, Japan has experienced deflation (declining prices) over this period. The reasons for this protracted period of stagnation continue to be debated by economists. Failure to recognize a change in the Japanese growth rate has hurt many investors, especially those taking a long-term perspective. From their peak in 1989, Japanese equity prices, as measured by the Nikkei index, fell by over 60 percent before bottoming out in mid-1992. Since that time, the market has been essentially flat despite considerable volatility.

The performance of the Japanese economy can be explained using the AD and AS model. The protracted slowdown of growth in Japan beginning in the early 1990s can be linked to the effect of the collapse of the equity and commercial real estate markets in the late 1980s and to excessive investment in capital goods (new factories and equipment) in the 1980s. These problems were compounded by persistent weakness in the banking sector, a profound lack of confidence among businesses and consumers, and negative demographics with slow growth in the working age population.

The sum of these developments caused a decline in both the AD and AS curves. Aggregate demand declined, causing the AD curve to shift to the left for the following reasons:

- The wealth effect due to the decline in equity and real estate prices sharply reduced consumption spending. Asset prices have yet to recover from the fall.
- Excessive investment in capital goods caused a sharp decline in business investment.
- There was a lack of confidence among businesses and consumers.
- Problems in the banking sector made monetary policy ineffective because banks were unable to lend, which negatively affected both consumer and business spending.

AS declined for the following reasons:

- Marked slowing in private investment spending reduced the capital stock. This also reduced potential GDP.
- Slow population growth limited the growth in the labor supply. This also reduced potential GDP.
- Higher energy prices slowed growth because of Japan's heavy dependence on imported energy.

As would be expected, the declines in both AD and AS resulted in slow GDP growth. The fact that prices fell indicates that the AD curve shifted more than the AS curve.

4. ECONOMIC GROWTH AND SUSTAINABILITY

We now shift focus from the short-run cyclical movement of the economy to its long-term growth rate. Economic growth is calculated as the annual percentage change in real GDP or the annual change in real per capita GDP:

- Growth in real GDP measures how rapidly the total economy is expanding.
- Per capita GDP, defined as real GDP divided by population, determines the standard of living in each country and the ability of the average person to buy goods and services.

Economic growth is important because rapid growth in per capita real GDP can transform a poor nation into a wealthy one. Even small differences in the growth rate of per capita GDP, if sustained over time, have a large impact on an economy's standard of living. One

should think of the growth rate of GDP as the equivalent of a rate of return on a portfolio. Small differences in return compounded over many years make a big difference. Nevertheless, there is a limit to how fast an economy can grow. Faster growth is not always better for an economy, because there are costs associated with excess growth, such as higher inflation, potential environmental damage, and the lower consumption and higher savings needed to finance the growth.

This raises the issue of sustainable growth, which requires an understanding of the concept of potential GDP. Recall that potential GDP measures the productive capacity of the economy and is the level of real GDP that an economy could produce if capital and labor are fully employed. In order to grow over time, an economy must add to its productive capacity. Thus, the **sustainable rate of economic growth** is measured by the rate of increase in the economy's productive capacity or potential GDP. It is important to note that economists cannot directly measure potential output. Instead, they estimate it using a variety of techniques discussed later in this chapter.

For global investors, estimating the sustainable rate of economic growth for an economy is important for both asset allocation and security selection decisions. Investors need to understand how the rate of economic growth differs among countries and whether these growth rates are sustainable. When examining the GDP data, global investors need to address a number of questions, including these three:

1. What are the underlying determinants or sources of growth for the country?
2. Are these sources of growth likely to remain stable or change over time?
3. How can we measure and forecast sustainable growth for different countries?

4.1. The Production Function and Potential GDP

The neoclassical or Solow growth model is the framework used to determine the underlying sources of growth for an economy. The model shows that the economy's productive capacity and potential GDP increase for two reasons:

1. Accumulation of such inputs as capital, labor, and raw materials used in production.
2. Discovery and application of new technologies that make the inputs in the production process more productive—that is, able to produce more goods and services for the same amount of input.

The model is based on a **production function** that provides the quantitative link between the level of output that the economy can produce and the inputs used in the production process, given the state of technology. A two-factor production function with labor and capital as the inputs is expressed mathematically as:

$$Y = AF(L, K)$$

where Y denotes the level of aggregate output in the economy, L is the quantity of labor or number of workers in the economy, K is the capital stock or the equipment and structures used to produce goods and services, and A represents technological knowledge or **total factor productivity** (TFP). TFP is a scale factor that reflects the portion of growth that is not accounted for by the capital and labor inputs. The main factor influencing TFP is technological change. Like potential GDP, TFP is not directly observed in the economy and must be estimated.

The production function shows that output in the economy depends on inputs and the level of technology. The economy's capacity to produce goods grows when these inputs increase and/or technology advances. The more technologically advanced an economy is, the more output it is able to produce from a given amount of inputs.

Two assumptions about the production function provide a link to microeconomics. First, we assume that the production function has constant returns to scale. This means that if all the inputs in the production process are increased by the same percentage, then output will rise by that percentage. Thus, doubling all inputs would double output. Second, we assume that the production function exhibits **diminishing marginal productivity** with respect to any individual input. This property plays an important role in assessing the contribution of labor and capital to economic growth. Marginal productivity looks at the extra output that is produced from a one-unit increase in an input if the other inputs are unchanged. It applies to any input as long as the other inputs are held constant. For example, if we have a factory of a fixed size and we add more workers to the factory, the marginal productivity of labor measures how much additional output each additional worker will produce.

Diminishing marginal productivity means that at some point the extra output obtained from each additional unit of the input will decline. In the preceding example, if we hire more workers at the existing factory (fixed capital input in this case), output will rise by a smaller and smaller amount with each additional worker. Traditionally, economists focused on the labor input and how the productivity of labor would decline given a fixed amount of land. The traditional growth theory, where labor is the only (variable) input, was developed by Thomas Malthus in his 1798 publication, *Essay on the Principle of Population*. Malthus argued that as the population and labor force grew, the additional output produced by an additional worker would decline essentially to zero and there would be no long-term economic growth. This gloomy forecast caused others to label economics the "dismal science."

The dire prediction implied by declining marginal productivity of labor never materialized, and economists changed the focus of the analysis away from labor to capital. In this case, if we add more and more capital to a fixed number of workers, the amount of additional output contributed by each additional amount of capital will fall. Thus, if capital grows faster than labor, capital will become less productive, resulting in slower and slower growth. Diminishing marginal productivity of capital has two major implications for potential GDP:

1. Long-term sustainable growth cannot rely solely on **capital deepening investment** that increases the stock of capital relative to labor. More generally, increasing the supply of some input(s) relative to other inputs will lead to diminishing returns and cannot be the basis for sustainable growth.
2. Given the relative scarcity and hence high productivity of capital in developing countries, the growth rates of developing countries should exceed those of developed countries. As a result, there should be a **convergence** of incomes between developed and developing countries over time.

Because of diminishing returns to capital, the only way to sustain growth in potential GDP per capita is through technological change or growth in total factor productivity. This results in an upward shift in the production function: The economy produces more goods and services using the same level of labor and capital inputs. In terms of the formal production function shown earlier, this is reflected by an increase in the technology parameter, A.

Using the production function, Robert Solow developed a model that explained the contribution of labor, capital, and technology (total factor productivity) to economic growth.

The growth accounting equation shows that the rate of growth of potential output equals growth in technology plus the weighted average growth rate of labor and capital.

$$\text{Growth in potential GDP} = \text{Growth in technology} + W_L(\text{Growth in labor}) + W_C(\text{Growth in capital})$$

where W_L and W_C are the relative shares of capital and labor in national income. The capital share is the sum of corporate profits, net interest income, net rental income, and depreciation divided by GDP. The labor share is employee compensation divided by GDP. For the United States, W_L and W_C are roughly 0.7 and 0.3, respectively.

The growth accounting equation highlights a key point: The contribution of labor and capital to long-term growth depends on their respective shares of national income. For the United States, because labor's share is higher, an increase in the growth rate of labor will have a significantly larger impact (roughly double) on potential GDP growth than will an equivalent increase in the growth rate of capital.

The growth accounting equation can be further modified to explain growth in per capita GDP. Because it measures the standard of living and purchasing power of the average person in an economy, per capita GDP is more relevant than the absolute level of GDP in comparing economic performance among countries. Transforming the growth accounting equation into per capita terms results in the following equation:

$$\text{Growth in per capita potential GDP} = \text{Growth in technology} + W_C(\text{Growth in capital-to-labor ratio})$$

The capital-to-labor ratio measures the amount of capital available per worker and is weighted by the share of capital in national income. Because capital's share in national income in the U.S. economy is 0.3, a 1 percent increase in the amount of capital available for each worker increases per capita output by only 0.3 percent. The equation shows that improvements in technology are more important than capital in raising an economy's standard of living.

4.2. Sources of Economic Growth

The growth accounting equation focuses on the main determinants of growth—capital, labor, and technology—and omits a number of other sources of growth to simplify the analysis. For many countries, however, natural resource and human capital inputs play an important role in explaining economic growth. Therefore, there are five important sources of growth for an economy:

1. Labor supply
2. Human capital
3. Physical capital
4. Technology
5. Natural resources

These sources of growth determine the capacity of the economy to supply goods and services.

4.2.1. Labor Supply

Growth in the number of people available for work (quantity of workforce) is an important source of economic growth and partially accounts for the superior growth performance, among the advanced economies, of the U.S. economy versus the European and Japanese economies. Most developing countries, such as China, India, and Mexico, have a large potential labor supply. We can measure the potential size of the labor input as the total number of hours available for work, which is given by:

$$\text{Total hours worked} = \text{Labor force} \times \text{Average hours worked per worker}$$

The **labor force** is defined as the portion of the working age population (over the age of 16) that is employed or available for work but unemployed. The contribution of labor to overall output is also affected by changes in the average hours worked per worker. Average hours worked is highly sensitive to the business cycle. However, the long-term trend has been toward a shorter workweek in the advanced countries. This development is the result of legislation, collective bargaining agreements, and the growth of part-time and temporary work.

4.2.2. Human Capital

In addition to the quantity of labor, the quality of the labor force is important. Human capital is the accumulated knowledge and skill that workers acquire from education, training, and life experience. It measures the quality of the workforce. In general, better-educated and skilled workers will be more productive and more adaptable to changes in technology.

An economy's human capital is increased through investment in education and on-the-job training. Like physical capital, investment in education is costly. Studies show that there is a significant return on education. That is, people with more education earn higher wages. Moreover, education may also have a spillover or externality impact: Increasing the educational level of one person not only raises the output of that person but also raises the output of those around him or her. The spillover effect operates through the link between education and advances in technology. Education not only improves the quality of the labor force but also encourages growth through innovation. Investment in health is also a major contributor to human capital, especially in developing countries.

4.2.3. Physical Capital Stock

The physical **capital stock** (accumulated amount of buildings, machinery, and equipment used to produce goods and services) increases from year to year as long as net investment (gross investment less depreciation of capital) is positive. Thus, countries with a higher rate of investment should have a growing physical capital stock and a higher rate of GDP growth. Exhibit 5-27 shows the level of business investment as a share of GDP. The exhibit shows significant variation across countries, with the investment share in the United States being low in comparison to other developed countries.

As is evident in Exhibit 5-27, the correlation between economic growth and investment is high. Countries that devote a large share of GDP to investment, such as China, India, and South Korea, have high growth rates. The fastest-growing countries in Europe over the past decade, Ireland and Spain, have the highest investment-to-GDP ratios. Countries that devote a smaller share of GDP to investment, such as Brazil and Mexico, have slower growth rates. The data show why the Chinese economy has expanded at such a rapid rate, achieving an

EXHIBIT 5-27 Business Investment as a Percentage of GDP

	1994	2000	2005	2007	Average Annual Real GDP Growth, 1991–2009
Developed Economies					
United States	17.2%	19.9%	19.2%	18.4%	2.2%
Japan	28.5	25.2	23.3	23.2	1.1
Germany	22.6	21.5	17.4	18.7	1.4
France	18.4	19.5	20.0	21.5	1.5
Italy	18.5	20.3	20.7	21.1	1.0
United Kingdom	16.1	17.1	16.9	17.8	2.2
Canada	18.8	19.2	21.3	22.6	2.1
Ireland	16.1	23.1	26.6	26.3	5.1
Spain	20.7	25.8	29.4	31.0	2.6
Australia	23.9	22.0	27.0	27.7	3.2
South Korea	36.4	31.1	29.3	28.8	4.9
New Zealand	20.9	20.4	24.1	22.9	2.7
Developing Countries					
Brazil	18.5%	16.8%	15.9%	17.5%	2.8%
China	34.5	34.3	41.0	40.0	10.2
India	NA	22.9	30.4	33.8	6.4
Indonesia	24.8	19.9	23.6	24.9	4.6
Mexico	19.4	21.4	20.1	20.8	2.4
Turkey	22.9	20.4	21.0	21.5	3.4

Source: OECD StatLink.

annual GDP growth rate of over 10 percent over the past two decades. Investment spending in China on new factories, equipment, and infrastructure as a percentage of GDP is the highest in the world. In recent years, China devoted over 40 percent of its GDP to investment spending.

4.2.4. Technology

The most important factor affecting economic growth is technology, especially in developed countries such as the United States. **Technology** refers to the process a company uses to transform inputs into outputs. Technological advances are discoveries that make it possible to produce more or higher-quality goods and services with the same resources or inputs. At the same time, technological progress results in the creation of new goods and services. Finally, technological progress improves how efficiently businesses are organized and managed.

Technological advances are very important because they allow an economy to overcome the limits imposed by diminishing marginal returns. Thus, an economy will face limits to growth if it relies exclusively on expanding the inputs or factors of production.

Because most technological change is embodied in new machinery, equipment, and software, physical capital must be replaced, and perhaps expanded, in order to take advantage of changes in technology. One of the key drivers of growth in developed countries over the past decade has been the information technology (IT) sector. Growth in the IT sector has been driven by technological innovation that has caused the prices of key technologies, such as semiconductors, to fall dramatically. The steep declines in prices have encouraged investment in IT at the expense of other assets. The sector has grown very fast and has made a significant contribution to economic growth, employment, and exports.

Countries can innovate through expenditures, both public and private, on research and development (R&D). Thus, expenditures on R&D and the number of patents issued, although not directly measuring innovation, provide some useful insight into innovative performance. Countries can also acquire new technology through imitation or copying the technology developed elsewhere. The embodiment of technology in capital goods can also enable relatively poor countries to jump ahead of the technology leaders.

Total factor productivity (TFP) is the component of productivity that proxies technological progress and organizational innovation. TFP is the amount by which output would rise because of improvements in the production process. It is calculated as a residual, the difference between the growth rate of potential output and the weighted average growth rate of capital and labor. Specifically,

$$\begin{aligned}\text{TFP growth} = {} & \text{Growth in potential GDP} \\ & - [W_L(\text{Growth in labor}) + W_C(\text{Growth in capital})]\end{aligned}$$

4.2.5. Natural Resources

Raw materials are an essential input to growth and include everything from available land to oil to water. Historically, consumption of raw materials has increased as economies have grown. There are two categories of natural resources:

1. **Renewable resources** are those that can be replenished, such as a forest. For example, if a tree is cut, a seedling can be planted and a new forest harvested in the future.
2. **Nonrenewable resources** are finite resources that are depleted once they are consumed. Oil and coal are examples.

Natural resources account for some of the differences in growth among countries. Today, such countries as Brazil and Australia, as well as those in the Middle East, have relatively high per capita incomes because of their resource bases. Countries in the Middle East have large pools of oil. Brazil has an abundance of land suitable for large-scale agricultural production, making it a major exporter of coffee, soybeans, and beef.

Even though natural resources are important, they are not necessary for a country to achieve a high level of income provided it can acquire the requisite inputs through trade. Countries in eastern Asia, such as Japan and South Korea, have experienced rapid economic growth but own few natural resources.

China's Economic Growth

Chinese economic growth, as measured by real GDP, has averaged about 10 percent since the late 1970s. On the demand side, high rates of business investment spending (over 40 percent of GDP, as shown in Exhibit 5-27) and export growth have fueled growth. On the supply side, there have been three major sources of growth over the past few decades:

1. Rapid capital accumulation due to the high rate of investment.
2. Adoption of more advanced technology from developed countries.
3. A large increase in the nonfarm labor force as people have moved from the interior rural areas to the urban areas on the coast.

China has not needed to rely on innovating new technology because it has pursued an export-oriented strategy based on imitation and replication of foreign products. Domestic natural resources have also not been an important driver of growth; China is a net importer of natural resources.

China faces three problems going forward that are likely to slow its growth rate. First, it is quite likely that its workforce will begin to shrink within the next decade. As a result of its long-standing one-child policy, overall population growth is very low (0.6 percent per year), and the number of people over 65 is the fastest-growing segment of the population. Thus, it is likely that the workforce will shrink over time. The impact may, however, be mitigated by the ongoing shift of workers from rural to urban areas of the country. In addition, China is investing heavily in education to improve the quality of the workforce.

The second potential problem arises from reliance on very high levels of investment spending. The high rate of investment has raised concerns about excess capacity in many industries and low rates of return on new investments in general. Because of diminishing returns to capital, further increases in capital per worker (capital deepening investment) will generate progressively smaller increases in real GDP per worker. In essence, China may be beyond the point where exceptional real growth can be sustained by simply adopting existing technology and deploying more of it.

The third potential problem looms on the demand side. The Chinese household sector has an extremely high propensity to save—so high that despite the very high level of investment spending, China runs a substantial current account surplus. If investment spending declines as a share of GDP, China will be left with a serious deficiency of aggregate demand unless household consumption, government expenditure, or both expand to fill the void.

In summary, China faces significant growth challenges. Decades of a policy designed to limit population growth may shift labor force growth into reverse just when the potency of rapid capital accumulation wanes. Meanwhile, even if very high levels of investment spending will not sustain historical levels of aggregate *supply* growth, any significant reduction in investment spending could be devastating to aggregate *demand* without a corresponding acceleration in household consumption.

4.3. Measures of Sustainable Growth

Measuring how fast an economy can grow is an important exercise. Economists project potential GDP into the future to forecast the sustainable growth path for the economy. An economy's potential GDP is an unobserved concept that is approximated using a number of alternative methods. It is important to note that estimates of the economy's potential growth can change as new data become available. Being able to understand such a change is critical for financial analysts because equity returns are highly dependent on the sustainable rate of economic growth.

We discussed in the previous section that the growth rate of potential GDP depends on the rate of technological progress as well as the growth rate of the labor force, physical and human capital, and natural resources.

How can we summarize all of these forces driving economic growth and develop a method to measure or estimate the growth rate of potential GDP? One way is to use the growth accounting equation discussed in Section 4.1.

$$\begin{aligned}\text{Growth in potential GDP} = {} & \text{Growth in technology} + W_L(\text{Growth in labor}) \\ & + W_C(\text{Growth in capital})\end{aligned}$$

The problem with this approach is that there are no observed data on potential GDP or on total factor productivity, and both must be estimated. In addition, data on the capital stock and the labor and capital shares of national income are not available for many countries, especially the developing countries.

As an alternative, we can focus on the productivity of the labor force, where we generally have more reliable data. **Labor productivity** is defined as the quantity of goods and services (real GDP) that a worker can produce in one hour of work. Our standard of living improves if we produce more goods and services for each hour of work. Labor productivity is calculated as real GDP for a given year divided by the total number of hours worked in that year, counting all workers. We use total hours, rather than the number of workers, to adjust for the fact that not everyone works the same number of hours.

$$\text{Labor productivity} = \text{Real GDP}/\text{Aggregate hours}$$

Therefore, we need to understand the forces that make labor more productive. Productivity is determined by the factors that we examined in the preceding section: education and skill of workers (human capital), investments in physical capital, and improvements in technology. An increase in any of these factors will increase the productivity of the labor force. The factors determining labor productivity can be derived from the production functions under the assumption of constant returns to scale, where a doubling of inputs causes outputs to double as well. Dividing the production function by $1/L$, we get the following:

$$Y/L = AF(1, K/L)$$

where Y/L is output per worker, which is a measure of labor productivity. The equation states that labor productivity depends on physical capital per worker (K/L) and technology (A). Recall that A can also be interpreted as total factor productivity. As this equation indicates, labor productivity and total factor productivity are related but distinct concepts. TFP is a scale factor that does not depend on the mix of inputs. Changes in TFP are measured as a residual,

capturing growth that cannot be attributed to specific inputs. In contrast, as shown in this equation, labor productivity—output per worker—depends on both the general level of productivity (reflected in TFP) and the mix of inputs. Increases in either TFP or the capital-to-labor ratio boost labor productivity. Because both output and labor input can be observed, labor productivity can be measured directly.

Labor productivity is a key concept for measuring the health and prosperity of an economy and its sustainable rate of growth. An analyst examining the growth prospects for an economy needs to focus on the labor productivity data for that country. Labor productivity largely explains the differences in the living standards and long-term sustainable growth rates among countries. The distinction between the level and growth rate of productivity is important to understand. Exhibit 5-28 provides such a comparison for selected countries.

4.3.1. Level of Labor Productivity

The higher the level of labor productivity, the more goods and services the economy can produce with the same number of workers. The level of labor productivity depends on the accumulated stock of human and physical capital and is much higher in the developed countries. For example, China has a population of over 1.3 billion people, compared with slightly over 300 million people in the United States. Because of its much larger population, China has significantly more workers than the United States. The U.S. economy as measured by real GDP is much larger, however, because U.S. workers are much more productive than

EXHIBIT 5-28 Labor Productivity: Level versus Growth Rate in Select Countries

	Level of Labor Productivity (2008 GDP per hour worked)	Labor Productivity Average Annual Growth Rate, 2001–2008
United States	$55.3	2.0%
Ireland	54.7	2.4
France	53.2	1.3
Germany	50.5	1.5
Sweden	45.9	2.0
United Kingdom	44.9	2.1
Canada	43.2	0.9
Spain	42.5	0.9
Italy	41.1	0.0
Japan	38.3	1.9
Greece	32.1	2.2
South Korea	25.3	4.7
Turkey	23.8	NA
Mexico	18.6	0.5

Source: OECD.StatExtracts (http://stats.oecd.org).

Chinese workers. As shown in Exhibit 5-28, the United States has the highest level of productivity in the world, producing over $55 of GDP per hour worked. Similarly, workers in France, Germany, and Ireland have high levels of productivity. In comparison, Mexican workers produce only $18.6 worth of GDP per hour worked. Thus, U.S. workers are nearly three times more productive than Mexican workers.

4.3.2. Growth Rate of Labor Productivity

The growth rate of labor productivity is the percentage increase in productivity over a year. It is among the economic statistics that economists and financial analysts watch most closely. In contrast to the level of productivity, the growth rate of productivity is typically higher in the developing countries, where human and physical capital is scarce but growing rapidly.

If productivity growth is rapid, it means the same number of workers can produce more and more goods and services. In this case, companies can afford to pay higher wages and still make a profit. Thus, high rates of productivity growth will translate into rising profits and higher stock prices.

EXAMPLE 5-14 Prospects for Equity Returns in Mexico

John Todd, CFA, manages a global mutual fund with nearly 30 percent of its assets invested in Europe. Because of the low population growth rate, he is concerned about the long-term outlook for the European economies. With potentially slower economic growth in Europe, the environment for equities may be less attractive. Therefore, he is considering reallocating some of the assets from Europe to Mexico. Based on the data in Exhibits 5-27 and 5-28, do you think that investment opportunities are favorable in Mexico? According to the OECD, the Mexican population increased by 0.8 percent in 2008, compared with a 0.3 percent increase in the European Union (27 countries).

Solution: Other than the higher population growth rate, the potential sources of growth for Mexico are not favorable. The level of business investment (Exhibit 5-27) in Mexico is quite low, especially in comparison to China, and even below that of many of the advanced economies in Europe, such as Spain and Italy. The level of labor productivity in Mexico is well below the levels in most European countries. This is not surprising given that the amount of capital per worker in Mexico is much lower than that in Europe. What is surprising and of concern is the rate of labor productivity growth in Mexico. Labor productivity in Mexico is growing at only a 0.5 percent annual rate, well below the rates of Germany, France, and the United Kingdom. This means that the rightward shift in the AS curve is greater for the European countries than for Mexico, despite the more favorable demographic trend in Mexico. In addition, it implies that there is more potential for expanding profit margins in Europe than in Mexico. Thus, the analysis of potential growth does not suggest a favorable outlook for equity returns in Mexico. In the absence of more favorable considerations (e.g., compelling equity valuations), John Todd should decide not to reallocate assets from Europe to Mexico.

In contrast, persistently low productivity growth suggests the economy is in bad shape. Without productivity gains, businesses have to either cut wages or boost prices in order to increase profit margins. Low rates of productivity growth should be associated with slow growth in profits and flat or declining stock prices.

4.3.3. Measuring Sustainable Growth

Labor productivity data can be used to estimate the rate of sustainable growth of the economy. A useful way to describe potential GDP is as a combination of aggregate hours worked and the productivity of those workers:

$$\text{Potential GDP} = \text{Aggregate hours worked} \times \text{Labor productivity}$$

Transforming the equation into growth rates, we get the following:

$$\begin{aligned}\text{Potential growth rate} = {} & \text{Long-term growth rate of labor force} \\ & + \text{Long-term labor productivity growth rate}\end{aligned}$$

Thus, potential growth is a combination of the long-term growth rate of the labor force and the long-term growth rate of labor productivity. Therefore, if the labor force is growing at 1 percent per year and productivity per worker is rising at 2 percent per year, then potential GDP (adjusted for inflation) is rising at 3 percent per year.

EXAMPLE 5-15 Estimating the Rate of Growth in Potential GDP

Exhibit 5-29 provides data on sources of growth for Canada, Germany, Japan, and the United States. Estimate the growth rates of the labor force, labor productivity, and potential GDP for each country by averaging the growth rates for these variables for the past two decades.

EXHIBIT 5-29 Sources of Growth: Average Annual Growth Rate

	1971–1980	1981–1990	1991–2000	2001–2008
Canada				
Labor force	2.1%	1.8%	1.1%	1.5%
Productivity	1.8	1.0	1.8	0.9
GDP	4.0	2.8	2.9	2.4
Germany				
Labor force	–0.9%	0.0%	–0.4%	–0.4%

(*Continued*)

EXHIBIT 5-29 *Continued*

	1971–1980	1981–1990	1991–2000	2001–2008
Productivity	3.7	2.3	2.5	1.5
GDP	2.9	2.3	2.1	1.0
Japan				
Labor force	0.3%	0.5%	–0.9%	–0.7%
Productivity	4.2	3.4	2.2	2.1
GDP	4.5	3.9	1.2	1.4
United States				
Labor force	1.6%	1.8%	1.5%	0.3%
Productivity	1.6	1.4	1.8	2.0
GDP	3.2	3.2	3.3	2.2

Solution: Potential GDP is calculated as the sum of the trend growth rate in the labor force and the trend growth rate in labor productivity. The growth in the labor force can differ from the population growth rate because of changes in the labor force participation rate and changes in hours worked per person. Estimating based on the average for the past two decades gives:

	Projected Growth in Labor Force	Projected Growth in Labor Productivity	Projected Growth in Potential GDP
Canada	1.3%	1.3%	2.6%
Germany	–0.4	2.0	1.6
Japan	–0.8	2.1	1.3
United States	0.9	1.9	2.8

The most striking result is the difference in labor force growth in Germany and Japan in contrast to labor force growth in the United States and Canada. Most of the differences in the growth rates in potential GDP of these countries can be explained by the demographic factors. The results suggest that Japan's sluggish growth over the past two decades is likely to continue. The weak productivity growth in Canada is of concern and is indicative of a low rate of innovation among Canadian companies.

EXAMPLE 5-16 Prospects for Fixed-Income Investments

As a fixed-income analyst for a large Canadian bank, you have just received the latest GDP forecast from the OECD for Canada, Germany, Japan, and the United States. The forecast is given here:

EXHIBIT 5-30 OECD GDP Forecast

	Projected Average Annual GDP Growth, 2010–2012
Canada	4.0%
Germany	1.5
Japan	0.5
United States	3.8

To evaluate the future prospects for fixed-income investments, analysts must estimate the future rate of inflation and assess the possibility of changes in monetary policy by the central bank. An important indicator for both of these factors is the degree of slack in the economy. One way to measure the degree of slack in the economy is to compare the growth rates of actual GDP and potential GDP.

Based on the estimates of potential GDP from the previous example and the information in Exhibit 5-30, evaluate the prospects for fixed-income investments in each of the countries.

Solution: In comparing the OECD forecast for GDP growth with the estimated growth rate in potential GDP, there are two cases to consider:

1. If actual GDP is growing at a faster rate than potential GDP, it signals growing inflationary pressures and an increased likelihood that the central bank will raise interest rates.
2. If actual GDP is growing at a slower rate than potential GDP, it signals growing resource slack, less inflationary pressures, and an increased likelihood that the central bank will reduce rates or leave them unchanged.

Exhibit 5-31 provides a comparison of actual and potential GDP for each of the countries.

The data suggest that inflationary pressure will grow in the United States and Canada and that both the U.S. Federal Reserve and the Bank of Canada will eventually raise interest rates. Thus, the environment for bond investing is not favorable in the United States and Canada, because bond prices are likely to decline.

EXHIBIT 5-31 Actual versus Potential GDP

	Projected Average Annual GDP Growth, 2010–2012	Potential GDP Growth
Canada	4.0%	2.6%
Germany	1.5	1.6
Japan	0.5	1.3
United States	3.8	2.8

With Germany growing at its potential rate of GDP growth, the rate of inflation should neither rise nor fall. Monetary policy is set by the European Central Bank (ECB), but data on the German economy play a big role in the ECB's decision. Based on the data, no change in ECB policy is likely. For bond investors, little change in bond prices is likely in Germany, so investors need to focus on the interest (coupon) income received from the bond.

Finally, growing resource slack in Japan will put downward pressure on inflation and may force the Bank of Japan to keep rates low. Bond prices should rise in this environment.

5. SUMMARY

This chapter has introduced important macroeconomic concepts and principles for macroeconomic forecasting and related investment decision making. Macroeconomics examines the economy as a whole by focusing on a country's aggregate output of final goods and services, total income, aggregate expenditures, and the general price level. The first step in macroeconomic analysis is to measure the size of an economy. Gross domestic product (GDP) enables us to assign a monetary value to an economy's level of output or aggregate expenditures. The interaction of aggregate demand and aggregate supply determines the level of GDP as well as the general price level. The business cycle reflects shifts in aggregate demand and short-run aggregate supply. The long-term sustainable growth rate of the economy depends on growth in the supply and quality of inputs (labor, capital, and natural resources) and advances in technology. From an investment perspective, macroeconomic analysis and forecasting are important because business profits, asset valuations, interest rates, and inflation rates depend on the business cycle in the short to intermediate term and on the drivers of sustainable economic growth in the long term. In addition, it is important to understand fiscal and monetary policies' economic impact on and implications for inflation, household consumption and saving, capital investment, and exports.

- GDP is the market value of all final goods and services produced within a country in a given time period.
- GDP can be valued by looking at either the total amount spent on goods and services produced in the economy or the income generated in producing those goods and services.

- GDP counts only final purchases of newly produced goods and services during the current time period. Transfer payments and capital gains are excluded from GDP.
- With the exception of owner-occupied housing and government services, which are estimated at imputed values, GDP includes only goods and services that are valued by being sold in the market.
- Intermediate goods are excluded from GDP in order to avoid double counting.
- GDP can be measured either from the value of final output or by summing the value added at each stage of the production and distribution process. The sum of the value added by each stage is equal to the final selling price of the good.
- Nominal GDP is the value of production using the prices of the current year. Real GDP measures production using the constant prices of a base year. The GDP deflator equals the ratio of nominal GDP to real GDP.
- Households earn income in exchange for providing—directly or indirectly through ownership of businesses—the factors of production (labor, capital, natural resources including land). From this income, they consume, save, and pay net taxes.
- Businesses produce most of the economy's output/income and invest to maintain and expand productive capacity. Companies retain some earnings but pay out most of their revenue as income to the household sector and as taxes to the government.
- The government sector collects taxes from households and businesses and purchases goods and services from the private business sector.
- Foreign trade consists of exports and imports. The difference between the two is net exports. If net exports are positive, then the country spends less than it earns; if exports are negative, it spends more than it earns. Net exports are balanced by accumulation of either claims on the rest of the world (net exports > 0) or obligations to the rest of the world (net exports < 0).
- Capital markets provide a link between saving and investment in the economy.
- From the expenditure side, GDP includes personal consumption (C), gross private domestic investment (I), government spending (G), and net exports ($X - M$).
- The major categories of expenditure are often broken down into subcategories. Gross private domestic investment includes both investment in fixed assets (plant and equipment) and the change in inventories. In some countries, government investment spending is separated from other government spending.
- National income is the income received by all factors of production used in the generation of final output. It equals GDP minus the capital consumption allowance (depreciation) and a statistical discrepancy.
- Personal income reflects pretax income received by households. It equals national income plus transfers minus undistributed corporate profits, corporate income taxes, and indirect business taxes.
- Personal disposable income equals personal income minus personal taxes.
- Private saving must equal investment plus the fiscal and trade deficits. That is, $S = I + (G - T) + (X - M)$.
- Consumption spending is a function of disposable income. The marginal propensity to consume represents the fraction of an additional unit of disposable income that is spent.
- Investment spending depends on the average interest rate and the level of aggregate income. Government purchases and tax policy are often considered to be exogenous variables determined outside the macroeconomic model. Actual taxes collected depend on income and are, therefore, endogenous—that is, determined within the model.
- The IS curve reflects combinations of GDP and the real interest rate such that aggregate income/output equals planned expenditures. The LM curve reflects combinations of GDP and the interest rate such that demand and supply of real money balances are equal.

- Combining the IS and LM relationships yields the aggregate demand curve.
- Aggregate demand and aggregate supply determine the level of real GDP and the price level.
- The aggregate demand curve is the relationship between real output (GDP) demanded and the price level, holding underlying factors constant. Movements along the aggregate demand curve reflect the impact of price on demand.
- The aggregate demand curve is downward sloping because a rise in the price level reduces wealth, raises real interest rates, and raises the price of domestically produced goods versus foreign goods. The aggregate demand curve is drawn assuming a constant money supply.
- The aggregate demand curve will shift if there is a change in a factor (other than price) that affects aggregate demand. These factors include household wealth, consumer and business expectations, capacity utilization, monetary policy, fiscal policy, exchange rates, and foreign GDP.
- The aggregate supply curve is the relationship between the quantity of real GDP supplied and the price level, keeping all other factors constant. Movements along the supply curve reflect the impact of price on supply.
- The short-run aggregate supply curve is upward sloping because higher prices result in higher profits and induce businesses to produce more and laborers to work more. In the short run, some prices are sticky, implying that some prices do not adjust to changes in demand.
- In the long run, all prices are assumed to be flexible. The long-run aggregate supply curve is vertical because input costs adjust to changes in output prices, leaving the optimal level of output unchanged. The position of the curve is determined by the economy's level of potential GDP.
- The level of potential output, also called the full employment or natural level of output, is unobservable and difficult to measure precisely. This concept represents an efficient and unconstrained level of production at which companies have enough spare capacity to avoid bottlenecks and there is a balance between the pool of unemployed workers and the pool of job openings.
- The long-run aggregate supply curve will shift because of changes in labor supply, supply of physical and human capital, and productivity/technology.
- The short-run supply curve will shift because of changes in potential GDP, nominal wages, input prices, expectations about future prices, business taxes and subsidies, and the exchange rate.
- The business cycle and short-term fluctuations in GDP are caused by shifts in aggregate demand and aggregate supply.
- When the level of GDP in the economy is below potential GDP, such a recessionary situation exerts downward pressure on the aggregate price level.
- When the level of GDP is above potential GDP, such an overheated situation puts upward pressure on the aggregate price level.
- Stagflation, a combination of high inflation and weak economic growth, is caused by a decline in short-run aggregate supply.
- The sustainable rate of economic growth is measured by the rate of increase in the economy's productive capacity or potential GDP.
- Growth in real GDP measures how rapidly the total economy is expanding. Per capita GDP, defined as real GDP divided by population, reflects the standard of living in a country. Real GDP growth rates and levels of per capita GDP vary widely among countries.
- The sources of economic growth include the supply of labor, the supply of physical and human capital, raw materials, and technological knowledge.

- Output can be described in terms of a production function. For example, $Y = AF(L, K)$ where L is the quantity of labor, K is the capital stock, and A represents technological knowledge or total factor productivity. The function $F(\bullet)$ is assumed to exhibit constant returns to scale but diminishing marginal productivity for each input individually.
- Total factor productivity is a scale factor that reflects the portion of output growth that is not accounted for by changes in the capital and labor inputs. TFP is mainly a reflection of technological change.
- Based on a two-factor production function, Potential GDP growth = Growth in TFP + W_L (Growth in labor) + W_C (Growth in capital), where W_L and W_C (= 1 − W_L) are the shares of labor and capital in GDP.
- Diminishing marginal productivity implies that:
  - Increasing the supply of some input(s) relative to other inputs will lead to diminishing returns and cannot be the basis for sustainable growth. In particular, long-term sustainable growth cannot rely solely on capital deepening—that is, increasing the stock of capital relative to labor.
  - Given the relative scarcity and hence high productivity of capital in developing countries, the growth rates of developing countries should exceed those of developed countries.
- The labor supply is determined by population growth, the labor force participation rate, and net immigration. The capital stock in a country increases with investment. Correlation between long-run economic growth and the rate of investment is high.
- In addition to labor, capital, and technology, human capital—essentially, the quality of the labor force—and natural resources are important determinants of output and growth.
- Technological advances are discoveries that make it possible to produce more and higher-quality goods and services with the same resources or inputs. Technology is the main factor affecting economic growth in developed countries.
- The sustainable rate of growth in an economy is determined by the growth rate of the labor supply plus the growth rate of labor productivity.

PRACTICE PROBLEMS[17]

1. Which of the following statements is the *most* appropriate description of gross domestic product (GDP)?
   A. GDP is the total income earned by all households, firms, and the government whose value can be verified.
   B. GDP is the total amount spent on all final goods and services produced within the economy over a given time period.
   C. GDP is the total market value of resalable and final goods and services produced within the economy over a given time period.

2. The component *least likely* to be included in a measurement of gross domestic product (GDP) is:
   A. the value of owner-occupied rent.
   B. the annual salary of a local police officer.
   C. environmental damage caused by production.

[17]These practice problems were written by Ryan C. Fuhrmann, CFA (Westfield, Indiana, USA).

3. Which of the following conditions is *least likely* to increase a country's GDP?
   A. An increase in net exports
   B. Increased investment in capital goods
   C. Increased government transfer payments

4. Which of the following would be included in Canadian GDP for a given year? The market value of:
   A. wine grown in Canada by U.S. citizens.
   B. electronics made in Japan and sold in Canada.
   C. movies produced outside Canada by Canadian filmmakers.

5. Suppose a painting is created and sold in 2010 for £5,000. The expenses involved in producing the painting amounted to £2,000. According to the sum-of-value-added method of calculating GDP, the value added by the final step of creating the painting was:
   A. £2,000.
   B. £3,000.
   C. £5,000.

6. A GDP deflator less than 1 indicates that an economy has experienced:
   A. inflation.
   B. deflation.
   C. stagflation.

7. The *most* accurate description of nominal GDP is:
   A. a measure of total expenditures at current prices.
   B. the value of goods and services at constant prices.
   C. a measure to compare one nation's economy to another.

8. From the beginning to the ending years of a decade, the annual value of final goods and services for country X increased from €100 billion to €300 billion. Over that time period, the GDP deflator increased from 111 to 200. Over the decade, real GDP for country X increased by approximately:
   A. 50 percent.
   B. 67 percent.
   C. 200 percent.

9. If the GDP deflator values for 2008 and 2010 were 190 and 212.8, respectively, which of the following *best* describes the annual growth rate of the overall price level?
   A. 5.8 percent
   B. 6 percent
   C. 12 percent

10. The numerator of the GDP price deflator reflects:
    A. the value of base-year output at current prices.
    B. the value of current-year output at current prices.
    C. the value of current-year output at base-year prices.

11. Consider the following data for 2010 for a hypothetical country:

Account Name	Amount ($ trillions)
Consumption	15.0
Statistical discrepancy	0.5
Capital consumption allowance	1.5
Government spending	3.8
Imports	1.7
Gross private domestic investment	4.0
Exports	1.5

Based only on the data given, the gross domestic product and national income are respectively *closest* to:
A. 21.1 and 20.6.
B. 22.6 and 20.6.
C. 22.8 and 20.8.

12. In calculating personal income for a given year, which of the following would *not* be subtracted from national income?
A. Indirect business taxes
B. Undistributed corporate profits
C. Unincorporated business net income

13. Equality between aggregate expenditure and aggregate output implies that the government's fiscal deficit must equal:
A. Private saving − Investment − Net exports.
B. Private saving − Investment + Net exports.
C. Investment − Private saving + Net exports.

14. Because of a sharp decline in real estate values, the household sector has increased the fraction of disposable income that it saves. If output and investment spending remain unchanged, which of the following is *most likely*?
A. A decrease in the government deficit
B. A decrease in net exports and increased capital inflow
C. An increase in net exports and increased capital outflow

15. Which curve represents combinations of income and the real interest rate at which planned expenditure equals income?
A. The IS curve
B. The LM curve
C. The aggregate demand curve

16. An increase in government spending would shift the:
    A. IS curve and the LM curve.
    B. IS curve and the aggregate demand curve.
    C. LM curve and the aggregate demand curve.

17. An increase in the nominal money supply would shift the:
    A. IS curve and the LM curve.
    B. IS curve and the aggregate demand curve.
    C. LM curve and the aggregate demand curve.

18. An increase in the price level would shift the:
    A. IS curve.
    B. LM curve.
    C. aggregate demand curve.

19. As the price level declines along the aggregate demand curve, the interest rate is *most likely* to:
    A. decline.
    B. increase.
    C. remain unchanged.

20. The full employment, or natural, level of output is *best* described as:
    A. the maximum level obtainable with existing resources.
    B. the level at which all available workers have jobs consistent with their skills.
    C. a level with a modest, stable pool of unemployed workers transitioning to new jobs.

21. Which of the following *best* describes the aggregate supply curve in the short run (e.g., one to two years)? The short run aggregate supply curve is:
    A. flat because output is more flexible than prices in the short run.
    B. vertical because wages and other input prices fully adjust to the price level.
    C. upward sloping because input prices do not fully adjust to the price level in the short run.

22. If wages were automatically adjusted for changes in the price level, the short-run aggregate supply curve would *most likely* be:
    A. more flat.
    B. steeper.
    C. unchanged.

23. The *least likely* cause of a decrease in aggregate demand is:
    A. higher taxes.
    B. a weak domestic currency.
    C. a fall in capacity utilization.

24. Which of the following is *most likely* to cause the long-run aggregate supply curve to shift to the left?
    A. Higher nominal wages
    B. A decline in productivity
    C. An increase in corporate taxes

25. Increased household wealth will *most likely* cause an increase in:
    A. household saving.
    B. investment expenditures.
    C. consumption expenditures.

26. The *most likely* outcome when both aggregate supply and aggregate demand increase is:
    A. a rise in inflation.
    B. higher employment.
    C. an increase in nominal GDP.

27. Which of the following is *least likely* to be caused by a shift in aggregate demand?
    A. Stagflation
    B. A recessionary gap
    C. An inflationary gap

28. Following a sharp increase in the price of energy, the overall price level is *most likely* to rise in the short run:
    A. and remain elevated indefinitely unless the central bank tightens.
    B. but be unchanged in the long run unless the money supply is increased.
    C. and continue to rise until all prices have increased by the same proportion.

29. Among developed economies, which of the following sources of economic growth is *most likely* to explain superior growth performance?
    A. Technology
    B. Capital stock
    C. Labor supply

30. Which of the following can be measured directly?
    A. Potential GDP
    B. Labor productivity
    C. Total factor productivity

31. The sustainable growth rate is *best* estimated as:
    A. the weighted average of capital and labor growth rates.
    B. growth in the labor force plus growth of labor productivity.
    C. growth in total factor productivity plus growth in the capital-to-labor ratio.

32. In the neoclassical or Solow growth model, an increase in total factor productivity reflects an increase in:
    A. returns to scale.
    B. output for given inputs.
    C. the sustainable growth rate.

The following information relates to Questions 33 and 34.

An economic forecasting firm has estimated the following equation from historical data based on the neoclassical growth model:

$$\text{Potential output growth} = 1.5 + 0.72(\text{Growth of labor}) + 0.28(\text{Growth of capital})$$

33. The intercept (1.5) in this equation is *best* interpreted as:
    A. the long-run sustainable growth rate.
    B. the growth rate of total factor productivity.
    C. above-trend historical growth that is unlikely to be sustained.

34. The coefficient on the growth rate of labor (0.72) in this equation is *best* interpreted as:
    A. the labor force participation rate.
    B. the marginal productivity of labor.
    C. the share of income earned by labor.

35. Convergence of incomes over time between emerging market countries and developed countries is *most likely* due to:
    A. total factor productivity.
    B. diminishing marginal productivity of capital.
    C. the exhaustion of nonrenewable resources.

CHAPTER 6

UNDERSTANDING BUSINESS CYCLES

Michele Gambera, CFA

Milton Ezrati

Bolong Cao, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Describe the business cycle and its phases.
- Explain the typical patterns of resource use fluctuation, housing sector activity, and external trade sector activity, as an economy moves through the business cycle.
- Describe theories of the business cycle.
- Describe types of unemployment and measures of unemployment.
- Explain inflation, hyperinflation, disinflation, and deflation.
- Explain the construction of indexes used to measure inflation.
- Compare inflation measures, including their uses and limitations.
- Distinguish between cost-push and demand-pull inflation.
- Describe economic indicators, including their uses and limitations.
- Identify the past, current, or expected future business cycle phase of an economy based on economic indicators.

1. INTRODUCTION

Agricultural societies experience good harvest years and bad ones. Weather is a main factor influencing whether the crops will be abundant or scarce, but of course there are also other factors, including plant and animal diseases. Ancient writings include stories of alternating good and bad production years. Modern diversified economies are less influenced by weather and diseases.

Whereas the chapter on national income accounting and growth focused on long-term economic growth and the factors that help foster it, this chapter addresses short-term movements in economic activity. Some of the factors causing such short-term movements are the same as those causing economic growth, such as changes in population, technology, and capital. However, other factors, such as money and inflation, are more specific to short-term fluctuations.

This chapter is organized as follows. Section 2 describes the business cycle and its phases, explaining the behaviors of businesses and households that typically characterize phases and transitions between phases. Section 3 provides an introduction to business cycle theory, in particular how different economic schools of thought interpret the business cycle and their recommendations with respect to it. Section 4 introduces basic concepts concerning unemployment and inflation, two important economic policy concerns that are sensitive to the business cycle. Section 5 discusses variables that fluctuate in predictable time relationships with the economy, focusing on variables whose movements have value in predicting the future course of the economy. A summary and practice problems conclude the chapter.

2. OVERVIEW OF THE BUSINESS CYCLE

Burns and Mitchell (1946) define the business cycle as follows:

> Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of events is recurrent but not periodic; in duration, business cycles vary from more than one year to 10 or 12 years.

This long definition is rich with important insights. First, business cycles are typical of economies that rely mainly on business enterprises—therefore, not agrarian societies or centrally planned economies. Second, a cycle has an expected sequence of phases representing alternation between expansion and contraction. Third, such phases occur at about the same time throughout the economy—that is, not just in agriculture or not just in tourism but in almost all sectors. Fourth, cycles are recurrent (i.e., they happen again and again over time) but not periodic (i.e., they do not all have the exact same intensity and duration). Finally, cycles typically last between one and 12 years.

Although Burns and Mitchell's definition may appear obvious in part, it indeed remains helpful even more than 60 years after it was written. Many investors like to think that there are simple regularities that occur at exactly the same time, every year or cycle: for example, shares always rally in January and big crashes occur in October. Of course, things are much more complex. The truth, as Burns and Mitchell remind us, is that history never repeats itself exactly, but it certainly has similarities that can be taken into account when analyzing the present and forecasting the future.

2.1. Phases of the Business Cycle

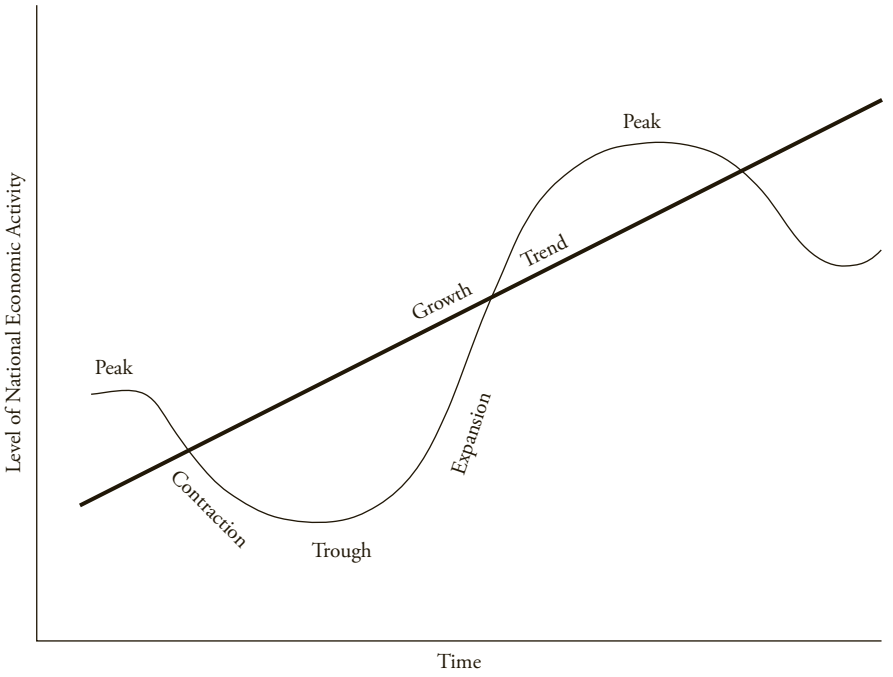
A business cycle consists of four phases: trough, expansion, peak, contraction. The period of **expansion** occurs after the **trough** (lowest point) of a business cycle and before its **peak**

(highest point), and **contraction** is the period after the peak and before the trough.[1] During the expansion phase, aggregate economic activity is increasing (*aggregate* is used because some individual economic sectors may not be growing). The contraction—often called a **recession**, but may be called a **depression** when exceptionally severe—is a period in which aggregate economic activity is declining (although some individual sectors may be growing). Business cycles are usually viewed as fluctuations around the trend growth of an economy, so such points as peaks and troughs are relative to the individual cycle.

Exhibit 6-1a shows a stylized representation of the business cycle, and Exhibit 6-1b provides a description of some important characteristics of each phase. The description distinguishes between early and late stages of the expansion phase, which are close to cycle turning points. Exhibit 6-1b also describes how several important economic variables evolve through the course of a business cycle.

The behavior of businesses and households frequently incorporates leads and lags, relative to what are established as turning points in a business cycle. For example, unemployment may not start decreasing immediately after the expansion phase starts because companies may want to fully use their existing workforces and wait to hire new employees until they are sure that the economy is indeed getting better. However, gradually all economic variables are going to revert toward their normal range of values (e.g., GDP growth will be a positive number). If any countercyclical economic policies were adopted during the recession, they would be gradually phased out; for example, if the central bank reduced interest rates to fight the

EXHIBIT 6-1a Schematic of Business Cycle Phases



[1]For more information, see www.nber.org/cycles/recessions.html.

EXHIBIT 6-1b Characteristics

	Early Expansion (Recovery)	Late Expansion	Peak	Contraction (Recession)
Economic activity	Gross domestic product (GDP), industrial production, and other measures of economic activity turn from decline to expansion.	Activity measures show an accelerating rate of growth.	Activity measures show a decelerating rate of growth.	Activity measures show outright declines.
Employment	Layoffs slow (and net employment turns positive), but new hiring does not yet occur and the unemployment rate remains high. At first, businesses turn to overtime and temporary employees to meet rising product demands.	Businesses begin full-time rehiring as overtime hours rise. The unemployment rate falls to low levels.	Businesses slow their rate of hiring; however, the unemployment rate continues to fall.	Businesses first cut hours and freeze hiring, followed by outright layoffs. The unemployment rate rises.
Consumer and business spending	Upturn is often most pronounced in housing, durable consumer items, and orders for light producer equipment.	Upturn becomes more broad-based. Businesses begin to order heavy equipment and engage in construction.	Capital spending expands rapidly, but the growth rate of spending starts to slow down.	Cutbacks appear most in industrial production, housing, consumer durable items, and orders for new business equipment, followed, with a lag, by cutbacks in other forms of capital spending.
Inflation	Inflation remains moderate and may continue to fall.	Inflation picks up modestly.	Inflation further accelerates.	Inflation decelerates but with a lag.

recession, it should start increasing rates toward their historical norms. Central banks are the monetary authority in most modern economies.

During a recession, investors place relatively high values on such safer assets as government securities and shares of companies with steady (or growing) positive cash flows, such as utilities and producers of staple goods. Such preferences reflect the fact that the marginal value of a safe income stream increases in periods when employment is insecure or declining. When asset markets expect the end of a recession and the beginning of an expansion phase, they will reprice risky assets upward. When an expansion is in sight, the markets will start incorporating higher profit expectations into the prices of corporate bonds and stocks, particularly those of such cyclical companies as producers of discretionary goods, for example automobiles. Typically, equity markets will hit the trough about three to six months before the economy and well before the economic indicators turn up. Indeed, the equity stock market is classified as a leading indicator of the economy.

When an economy's expansion is well established, a later part of an expansion called a **boom** often follows. The boom is still an expansionary phase, which is characterized by

economic growth testing the limits of the economy. For example, companies expand so much that they will have difficulty finding qualified workers, and therefore will start bidding wars to hire competitors' employees. The accompanying rise in labor costs causes a reduction of profits. Another example is that companies start believing that the economy will continue expanding for the foreseeable future, and therefore decide to borrow money to expand their production capacity. Clearly, these two examples represent situations that cannot go on forever; salaries and invested capital cannot grow exponentially without affecting profit margins. At this point, the government or its central bank often steps in to attempt to correct excesses. Consider the following situation. The central bank is concerned that excessive salary growth may lead to inflation. For example, companies will try to pass on higher production costs to their customers or excessive borrowing may cause investors to have cash flow problems. At the height of the boom phase, the economy is said to be overheating (just like the engine of a car that has been pushed to an excessive level).

During the boom, the riskiest assets will often have substantial price increases. Safe assets, such as government bonds that were more highly prized during the recession, may have lower prices and thus higher yields. In addition, investors may fear higher inflation, which also contributes to higher nominal yields.

The end of the expansion, or boom, is characterized by the peak of the business cycle, which is also the beginning of the contraction (also known as a downturn). Here, either because of restrictive economic policies established to tame an overheated economy or because of other shocks, such as energy prices or a credit crisis, the economy stumbles and starts slowing down. Unemployment will increase and GDP growth will decrease during this part of the business cycle.

EXAMPLE 6-1 When Do Recessions Begin and End?

A simple and commonly referred to rule is: A recession has started when a country or region experiences two consecutive quarters of negative real GDP growth. Real GDP growth is a measure of the inflation-adjusted growth of the overall economy. This rule can be misleading because it does not indicate a recession if real GDP growth is negative in one quarter, slightly positive the next quarter, and again negative in the next quarter. Many analysts question this result. This issue is why, in some countries, there are statistical and economic committees that apply the principles stated by Burns and Mitchell to several macroeconomic variables (and not just real GDP growth) as a basis to identify business cycle peaks and troughs. The National Bureau of Economic Research (NBER) is the well-known organization that dates business cycles in the United States. Interestingly, the economists and statisticians on NBER's Business Cycle Dating Committee analyze numerous time series of data focusing on employment, industrial production, and sales. Because the data are available with a delay (preliminary data releases can be revised even one year after the period they refer to), it also means that the NBER committee's determinations may take place well after the business cycle turning points have occurred. As we will see later in the chapter, there are practical indicators that may help economists understand in advance if a cyclical turning point is about to happen.

1. Which of the following rules is *most* commonly used to determine when recessions start? Recessions start when:
   A. the central bank runs out of foreign reserves.
   B. real GDP has two consecutive quarters of negative growth.
   C. economic activity experiences a material decline in two business sectors.
2. Suppose you are interested in forecasting earnings growth for a company active in a country where no official business cycle dating committee (such as at the NBER) exists. Which variables would you consider in estimating peaks and troughs of that country's business cycles, in addition to any existing index of leading indicators?
   A. Inflation, the central bank's discount rate, and unemployment.
   B. The Dow Jones Industrial Average (DJIA), equity market average book value, and monetary base.
   C. Unemployment, GDP growth, industrial production, and inflation.

Solution to 1: B is correct. GDP is the measure for the whole economy, whereas foreign reserves or a limited number of sectors may not have a material impact on the whole economy.

Solution to 2: C is correct. The discount rate, the monetary base, and stock market indexes are not direct measures of economic activities. The first two are determined by monetary policy, which reacts to economic activities, whereas the stock market indexes tend to be forward-looking.

The types of assets that tend to outperform during the contraction phase vary with the events that trigger the crisis. For example, in the case of a bank crisis, investors may panic and try to sell risky assets at any price, hoping to buy safer assets. In the case of a slowdown as a result of restrictive economic policies within the country, investors may try to buy shares of exporting companies.

Investors, who are often optimists in the expansion phase, tend to be overly pessimistic at the bottom of the recession. During recent economic history in many countries, such as the United States, economic contractions have been shorter than expansions.

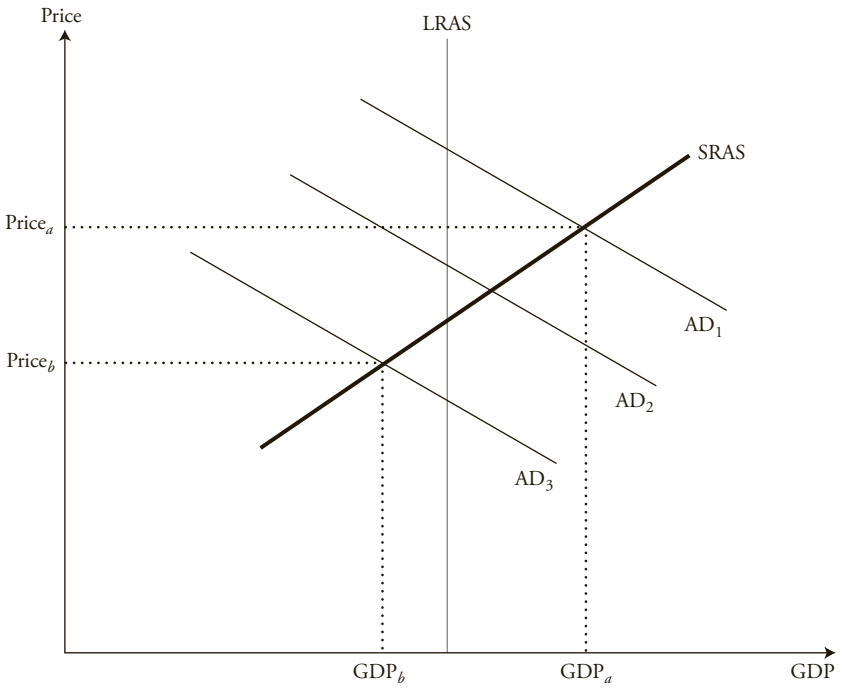
Many economic variables and sectors of the economy have distinctive cyclical patterns. Knowledge of these patterns can offer insight into likely cyclical directions overall or can be particularly applicable to an investment strategy that requires more specific rather than general cyclical insights for investment success. The following sections provide overviews of how the use of resources (the factors of production) typically evolves through the business cycle and how the sectors of real estate and external trade characteristically behave.

2.2. Resource Use through the Business Cycle

This section provides a broad overview of how the use of resources needed to produce goods and services typically evolves during a business cycle.

There are significant links between fluctuations in inventory, employment, and investment in physical capital with economic fluctuations. When a downturn starts, for example because of a monetary or fiscal tightening, aggregate demand (AD) decreases, shifting the AD

EXHIBIT 6-2 Policy-Triggered Recession



curve left; as a result, inventories start to accumulate. Exhibit 6-2 displays this shift from AD_1 to AD_2 along with the long-run aggregate supply (LRAS) and short-run aggregate supply (SRAS) curves. This shift occurs because demand decreases and companies end up with an excess of inputs and intermediate products ready for production, as well as final products ready for sale. As a consequence, companies may slow down production, thus idling workers (e.g., no more overtime) and physical capital (e.g., the equipment is used at less than full capacity), which further decreases aggregate demand and shifts the AD curve even further to the left (in Exhibit 6-2, AD_2 shifts to AD_3). In total, the equilibrium has moved to lower prices (from $Price_a$ to $Price_b$) and lower GDP (from GDP_a to GDP_b).

Companies do not start firing workers right away. First of all, if it is just a temporary slowdown for the economy, these workers may be needed again soon, so it is better to retain their jobs. In particular, selecting and training new workers is costly, so it is efficient to keep workers on payroll, even if they are not fully utilized, while waiting out a short period of slow business. Second, some economists suggest that there is an implicit bond of loyalty between a company and its workers, and thus workers will be more productive if they know that the company is not disposing of them at the first sign of economic trouble.

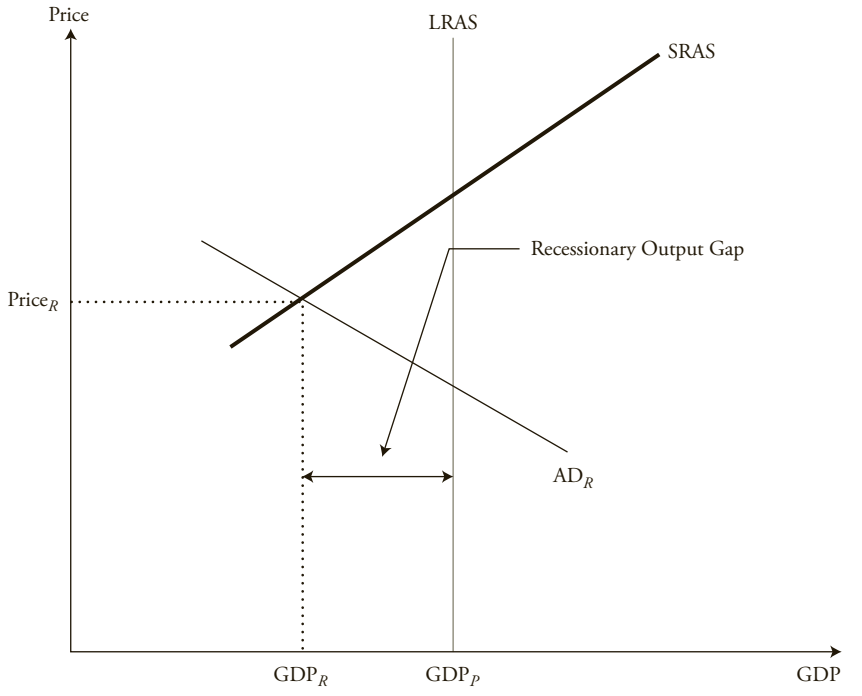
Even though companies may not fire workers right away, they will likely reduce production and stop ordering new inventories and new production equipment. For example, if business slows down, there is no need to order an additional delivery truck—the existing trucks will be sufficient because there are fewer deliveries to make. If enough companies act similarly, the slow patch in business will be exacerbated. By the same token, if workers start

worrying that lean years are coming, they may start saving more so they have more reserves. These actions will also reduce AD and further slow down the economy.

If the downturn becomes more severe, companies will start switching into recession mode—that is, they will cut all nonessential costs. This step often means terminating consultants, workers beyond the strict minimum, any standing supply orders, advertising campaigns, and so on. Capacity utilization will be low, and few companies will invest in new equipment and structures. In addition, because of the gloomy economic framework, banks will be wary of lending because bankruptcy risks will increase. Companies will try to liquidate their inventories of unsold products. As a result, the AD curve has a dramatic shift left toward an even lower GDP, and the economy enters what seems to be a downward spiral.

In an economic downturn, companies will probably not sell physical capital. First, it is difficult to find buyers. Second, physical capital becomes obsolete simply as a function of time. For example, the delivery truck mentioned earlier will depreciate even if it is being used at less than full capacity. Therefore, although the stock of inventories can adjust with sales prices set below the cost of production (fire sales), physical capital will adjust simply by way of aging plus technical obsolescence. Also, as mentioned earlier, if salaries do not fall immediately, the excess stock of labor will adjust via an increase in unemployment, as shown in Exhibit 6-3. The economy will be at a point where the short-run equilibrium ($Price_R$, GDP_R) is not on the LRAS curve. The gap between the recession output (GDP_R) and the potential output (GDP_P), the level of real GDP that could be achieved if all resources were fully employed, represents the level of slack resource (unemployment for labor and idleness

EXHIBIT 6-3 Recessionary Output Gap



for capital). Decreases in aggregate demand are likely to depress wages or wage growth as well as prices of inputs and capital goods. After a while, all of these input prices will be very low relative to their normal levels. In addition, the monetary authority may have cut interest rates to try to revive the economy.

As the price or inflation rate starts to fall, consumers and companies begin to purchase more—that is, aggregate quantity demanded of output begins to rise. (The short-run equilibrium moves downward along the AD curve.) A consequence of these economic influences is that some companies will increase production because of inventories being low. Also, because interest rates have been cut lower, some companies and households will determine that building a more efficient warehouse or renovating a home has become relatively cheap, and therefore will decide to start investing in structures, housing, and durable goods (equipment for companies, appliances for households). This stage is the turning point of the business cycle, when aggregate demand gradually starts to increase (the aggregate demand curve shifts right).

When the economic revival begins, companies will not immediately start the costly process of selecting and hiring new workers. They will wait for the expansion to give clear signs of life. However, if enough new investment triggers an increase in aggregate demand, companies will start replenishing their inventories. Low inventories mean that even if companies buy a little bit of inventory, there will be a short-term boom in demand for inputs and intermediate products, which will further support aggregate demand. This stage is often called inventory rebuilding or restocking in the financial press and may be followed by further capital expenditure (the one truck is now not sufficient for all deliveries). Demand for all factors of production increases.

As aggregate demand continues to grow, the boom phase of the cycle begins. Two different results of a boom phase could occur. First, the economy may experience shortages, and the demand for factors of production may exceed supply. Another possibility is that the excess demand comes from an overly optimistic buildup in production capacity, which means that supply of capital will greatly exceed demand a few months down the road (think of the excessive amount of fiber-optic investments during the 1990s technology boom or the residential overbuilding in many countries during the early 2000s housing bubble). These are possible triggers for the next recession.

2.2.1. Fluctuation in Capital Spending

This section describes how capital spending typically fluctuates with the business cycle. Because business profits and cash flows are extremely sensitive to the pace of economic activity, spending on new equipment and commercial structures is sensitive. Shifts in capital spending tend to affect the overall economic cycle in three stages or phases.

First, the downturn in spending on equipment usually occurs abruptly as final demand starts to fall off. Businesses, seeing a decline in sales and expecting a drop in profits and free cash flow, will halt new ordering and cancel existing orders if possible because they have no reason to expand production capacity in such a situation. The initial cuts occur in orders for technology and light equipment because there are short lead times from order to delivery, and cutbacks simply require managers to decide against any additional orders. Because it sometimes takes longer to cancel or even halt construction activity or the installation of larger, more complex pieces of equipment, cutbacks in these areas unfold with a longer lag. Typically, the initial cutbacks during this phase exaggerate the economy's downward thrust. Then later, as the general cyclical downturn matures, cutbacks in spending on structures and heavy equipment add to the negative cyclical momentum.

In the second phase, when the economy begins its initial recovery, sales are still at such low levels that a business hardly utilizes all its existing capacity and has little need to expand it. But although capacity utilization remains low, orders begin to pick up tentatively for two reasons. First, the economic improvement creates growth in earnings and free cash flow that gives businesses the financial wherewithal to increase spending. Second, the upturn in sales convinces managers to reinstate the more radical cancellations made in the uncertainty of the initial cyclical weakness, typically in equipment with a high rate of obsolescence, such as software, systems, and technological hardware. The shock of profit declines during a recession also instills a more general desire to make purchases that enhance efficiency more than those that expand capacity. As in the downturn, movements in new orders provide the first signal of recovery.

The third phase develops much later in the cyclical upturn, after a long period of output growth begins to strain the economy's overall productive capacity. Orders and sales at this phase of the cycle focus on capacity expansion and involve a higher proportion of heavy and complex equipment, warehouses, and factories. As the economic growth picks up speed, the capacity utilization starts to rise, a trend that cannot substitute for capital spending. In fact, spending on new capacity usually begins long before capacity would seem to need additions. This seeming disconnect occurs because economies are always changing their needs. Much that counts as capacity in the statistics becomes less relevant even though the underlying assets remain fully serviceable. The composition of the economy's capacity may not be optimal for the current economic structure, necessitating spending for new capital. A company, for instance, that needs more transportation equipment cannot substitute with a surplus of forklifts, although they are counted in overall capacity. Similarly, a company that needs warehouse space in the suburbs of Mumbai gains little relief from the surplus warehouse space in Goa. This last phase in the cyclical spending cycle may then occur surprisingly soon after capacity utilization picks up. Orders, of course, give the usual early indication of this phase.

The most watched indicator of the future direction of capital spending is orders for capital equipment, because they precede moves in actual shipments.

EXAMPLE 6-2 Capital Spending

1. The most likely reason that U.S. analysts often follow new orders for capital goods excluding defense and aircraft is because:
   A. the military is part of the public sector.
   B. aircraft and military orders are often the same, so there is double counting.
   C. armed forces and airlines tend to place infrequent and large (i.e., lumpy) orders, which create a false signal for the index.
2. Orders for equipment decline before construction orders in a recession because:
   A. businesses are uncertain about cyclical directions.
   B. they are easier to cancel than large construction contracts.
   C. businesses value light equipment less than structures and heavy machinery.

Solution to 1: C is correct. Business cycle indicators need to represent the activities in the whole economy and thus should not be influenced by some particular sectors that may have uncorrelated fluctuations.

Solution to 2: B is correct. Because it usually takes a much longer time to plan and complete large construction projects than for equipment orders, construction projects may be less influenced by business cycles.

Note: New orders statistics include orders that will be delivered over several years. For example, it is common for airlines to order 40 airplanes to be delivered over five years. Therefore, analysts use core orders that exclude defense and aircraft for a better understanding of the economy's trend.

2.2.2. Fluctuation in Inventory Levels

Inventory accumulation and decumulation by businesses can also occur with such rapidity and with such large movements that they have a much greater effect on economic growth than is justified by their relatively small aggregate size. The key indicator in this area is the inventory-to-sales ratio that measures the outstanding stock of available inventories relative to the level of sales. The interaction of this gauge with the cycle develops in three distinct stages.

First, toward the top of the economic cycle, as sales fall or slow, it usually takes businesses a while to cut back on new production. Inventories, as a consequence, accumulate involuntarily, which, combined with a fall in sales, prompts a sudden rise in inventory-to-sales ratios. Because these inventories count positively in the accounting for the whole economy, this rise falsely blunts signs of economic weakness, so practitioners look for figures that abstract from inventory swings that are commonly called "final sales." To adjust and sell off these unwanted inventories, a business has to cut production below even reduced sales levels. That action causes subsequent indicators in the overall economy to look weaker than they otherwise might have been. Although final sales offer a reality check, the production cutbacks involved in the inventory decumulations lead to order cancellations and layoffs that can subsequently cut final sales further and deepen cyclical corrections.

Second, with businesses producing at rates below the sales volumes necessary to dispose of unwanted inventories, inventory-to-sales ratios begin to fall back toward normal. When these indicators return to acceptable levels and businesses no longer have any need to reduce inventories further, they will raise production levels even without any sales growth just to end the decline in inventory levels. This step results in a seemingly improved economic situation, even if sales remain depressed. Again, final sales provide a reality check on the underlying economic situation. As a business reaches this phase in the cycle, the seemingly minor increase in production levels can actually mark the beginning of the cyclical turn because the relief of finally adjusting inventories can slow or stop the rate of staff layoffs and do the same for the business's other demands for inputs.

The third stage occurs as sales generally begin their cyclical upturn. In a manner analogous to but the opposite of the initial cyclical downturn, a business may initially fail to keep production on pace with sales, which causes it to lose inventory to the initial sales increase. The subsequent fall in inventory-to-sales ratios, when it occurs in the face of rising sales, quickly prompts a surge in production not only to catch up with sales but also to replenish depleted inventories. However, sometimes during short or severe recessions, when businesses have not had time to adjust or reduce inventories to acceptable levels, only the initial sales increase is necessary to adjust inventories, thereby making increased production unnecessary and extending the lag to the cyclical pickup in production. But whether the production upturn occurs with a short or a long lag, it typically marks the turn in hiring patterns and for a time can markedly exaggerate the cyclical strength.

EXAMPLE 6-3 Inventory Fluctuation

1. Though a small part of the overall economy, inventories can reflect growth significantly because they:
   A. reflect general business sentiment.
   B. tend to move forcefully up or down.
   C. determine the availability of goods for sale.
2. Inventories tend to rise when:
   A. inventory-to-sales ratios are low.
   B. inventory-to-sales ratios are high.
   C. economic activity begins to rebound.
3. Inventories will often fall early in a recovery because:
   A. businesses need profit.
   B. sales outstrip production.
   C. businesses ramp up production because of increased economic activity.

Solution to 1: B is correct. As stated in the chapter, inventory levels fluctuate dramatically over the business cycle.

Solution to 2: A is correct. When the economy starts to recover, sales of inventories can outpace production, which results in low inventory-to-sales ratios. Companies then need to accumulate more inventories to restore the ratio to a normal level.

Solution to 3: B is correct. The companies are slow to increase production in the early recovery phase because they first want to confirm that the recession is over. Increasing output also takes time after the downsizing during the recession.

2.2.3. Consumer Behavior

As the largest single sector of almost every developed economy (70 percent of the U.S. economy), patterns of household consumption determine overall economic direction more than any other sector. Patterns of consumption are critical to practitioners who, for any number of diverse reasons, have a particular interest in the sector. For example, marketers or equity analysts covering consumer product companies would have a high interest in the sector.

The two primary measures of household consumption are retail sales and, where available, a broad-based indicator of consumer spending that also includes purchases outside purely retail establishments, such as utilities, household services, and so on. Often these are presented in nominal terms and deflated to indicate directions of real or unit purchases and growth. Some indicators can make much finer distinctions, such as tracking spending, both real and nominal, of the more specific groups of consumer products. The three major divisions are: (1) durable goods, such as autos, appliances, and furniture; (2) nondurable goods, such as food, medicine, cosmetics, and clothing; and (3) services, such as medical treatment, entertainment, communications, and hairdressers. Because durable purchases usually replace items with longer useful lives, households in hard times can postpone such purchases more readily than spending on either services or nondurable goods. Comparing trends in durable purchases

to those in the other categories can give practitioners a notion of the economy's progress through the cycle, with weakness in durables spending an early indication of general economic weakness, and the catch-up in such spending a harbinger of a more general cyclical recovery.

Beyond direct observations of consumer spending and its mix, practitioners can also gauge future directions by analyzing consumer confidence or sentiment to ascertain how aggressive consumers may be in their spending. Usually, such information is in the form of surveys intended to provide practitioners with a general guide to trends. But in practice, they frequently do not reflect actual consumer behavior because survey respondents often answer what they imagine are the preferences of the typical consumer, indicating behavior contrary to their own.

Growth in income provides a more solid indication of consumption prospects, and household income figures are widely available in most countries. Especially relevant is after-tax income or, after deduction of necessary living expenses, what is frequently called disposable income. Some analysts chart consumer spending less from an examination of gross or even after-tax income than from a concept termed *permanent income*. Permanent income abstracts away from temporary income unsustainable losses or gains and tries to capture the income flow on which households believe they can rely. Spending on durables tends to rise and fall with the gains and losses whatever their cause, but the basic level of consumption reflects this notion of permanent income.

Even after making such distinctions, consumer spending frequently diverges from income, no matter how it is measured. An analysis of the saving rate can assist practitioners in this regard. Calculated in different ways in different countries and sometimes in different ways within the same country, cross-border comparisons of saving rates are difficult. But because all aim in one way or another to measure the percent of income households set aside from spending, variations in saving rates can capture consumers' willingness to reduce spending out of current income. The saving rate also reflects future income uncertainties perceived by consumers (precautionary savings). Therefore, it indicates consumers' ability to spend despite possible lower income in the future. Thus a rise in the saving rate, usually measured as a percentage of income, could indicate a certain caution among households and signal economic weakening. Certainly that was the case in Europe and the United States during the global financial crisis that led into the 2008–2009 recession. At the same time, the greater the stock of savings in the household sector and the wider the gap between ongoing income and spending, the greater the capability among households to pick up their spending, even ahead of income. So, although unusually high saving rates may at first say something negative about the cyclical outlook, they point longer-term to the potential for recovery.

EXAMPLE 6-4 Consumer Behavior

1. Durable goods have the most pronounced cyclical behavior because:
   A. they have longer useful lives.
   B. their purchase cannot be delayed.
   C. they are needed more than nondurable goods or services.
2. Permanent income provides a better guide to:
   A. saving rates.
   B. spending on services.
   C. spending on durable goods.

Solution to 1: A is correct. Durable goods are usually big-ticket items, the life span of which can be extended with repairs and without incurring high replacement costs. So consumers tend to delay replacement when the economic outlook is not favorable.

Solution to 2: B is correct. Households adjust consumption based on perceived permanent income level rather than temporary earning fluctuations. Saving rates and durable goods consumption are more related to the short-term uncertainties caused by recessions.

2.3. Housing Sector Behavior

Although generally a much smaller part of the overall economy than consumer spending, housing can move up and down so rapidly that it can count more in overall economic movements than the sector's relatively small size might suggest. Almost every major economy offers statistics on new and existing home sales, residential construction activity, and sometimes, importantly, the inventory of unsold homes on the market. Statistics are also potentially available for the average or median price of homes, sometimes recorded by type of housing unit and sometimes as the price per square foot or square meter. Whatever the specific statistics, the relationships in this area typically follow fairly regular cyclical patterns.

Because many home buyers finance their purchase with a mortgage, the sector is especially sensitive to interest rates. Home buying and consequently construction activity expand in response to lower mortgage rates and contract in response to higher mortgage rates.

Beyond such interest rate effects, housing also follows its own internal cycle. When housing prices are low relative to average incomes, and especially when mortgage rates are also low, the cost of owning a house falls and demand for housing increases. Often indicators of the cost of owning a house are available to compare household incomes with the cost of supporting an average house, both its price and the expense of a typical mortgage. Commonly, housing prices and mortgage rates rise disproportionately as expansionary cycles mature, bringing on an increase in relative housing costs, even as household incomes rise. The resulting slowdown of house sales can lead to a cyclical downturn first in buying and then, as the inventory of unsold houses builds, in actual construction activity.

These links, clear as they are, are far from mechanical. If housing prices have risen rapidly in the recent past, for instance, many people will buy to gain exposure to the expected price gains, even as the purchase in other respects becomes harder to rationalize. Such behavior can extend the cycle upward and may result in a more severe correction. This result occurs because "late buying" activity invites overbuilding. The large inventory of unsold homes eventually puts downward pressure on real estate prices, catching late buyers who have stretched their resources. This pattern occurred in many countries during the 2008–2009 global financial crisis.

Behind such cyclical considerations, housing, more than most economic sectors, responds to demographics, in particular the pace of family or household formation in an economy. Not every economy has data on family formation, but almost all offer information on the growth of particular age groups or cohorts in their respective populations. A focus on those cohorts, typically 25- to 40-year-olds, when household formation commonly occurs, usually can substitute for direct measures of net family formation. Adjusted for older people who are vacating existing homes, such calculations serve as an indicator of underlying, longer-term,

secular housing demand. Although such measures have little to do with business cycles, they do offer a gauge, along with affordability, of how quickly the housing market can correct excess and return to growth. In China, for instance, where the government recently estimated a need for about 400 million more urban housing units over the next 25 years, housing demand will more quickly reverse cyclical weakness than in such economies as Italy and Japan where net new family formation is relatively slight.

EXAMPLE 6-5 Housing Sector Behavior

1. Housing is more sensitive than other sectors of the economy to:
   A. interest rates.
   B. permanent income.
   C. government spending.
2. Apart from questions of affordability, house buying also reflects:
   A. the rate of family formation.
   B. speculation on housing prices.
   C. both the rate of family formation and speculation on housing prices.

Solution to 1: A is correct. Because real estate purchases are usually financed with mortgage loans, interest rate changes directly influence the monthly payment amounts.

Solution to 2: C is correct. Family formation constitutes the actual need for housing, whereas speculation on housing prices reflects the fact that real estate has investment value.

2.4. External Trade Sector Behavior

The external trade sector varies tremendously in size and importance from one economy to another. In such places as Singapore and Hong Kong, for instance, where almost all inputs are imported and the bulk of the economy's output finds its way to the export market, trade (counting exports and imports together) easily exceeds the GDP. In other places, such as the United States with its huge continental economy, external trade assumes a much smaller part of GDP. In recent decades, though, the relative size of international trade has grown in almost every country in the world. Because the external trade sector can be very important for some countries, the business cycles of the large open economies in the world can be transmitted to them through international trade.

Typically, imports rise, all else being equal, with the pace of domestic GDP growth, as needs and wants or generally rising demand also increase purchases of goods and services from abroad. Thus, imports respond to the domestic cycle. Exports are less a reflection of the domestic cycle than of cycles in the rest of the world. If these external cycles are strong, all else being equal, exports will grow even if the domestic economy should experience a decline in growth. To understand the impact of exports, financial analysts need to understand the strength of the major trading partners of the economy under consideration. The net effect of trade offsets cyclical weakness, and depending on the importance of exports to the economy, could erase it altogether. Most practitioners look at the net difference between exports and

imports (they use the balance of payments, which calculates trade's contribution to the economy as exports less imports). For these reasons, such differences can mean that the pattern of external trade balances is entirely different from the rest of the domestic economic cycle.

Currency also has an independent effect that can move trade in directions strikingly different from the domestic economic cycle. When a nation's currency appreciates, foreign goods seem cheaper than domestic goods to the domestic population, prompting, all else being equal, a relative rise in imports. At the same time, such currency appreciation makes that nation's exports more expensive on global markets, impairing their prospects. Of course, currency depreciation has the opposite effects. Although currency moves are sometimes violent, they have a significant effect on trade and the balance of payments only when they cumulate in a single direction for some time. Moves from one month or quarter to the next, however great, have a minimal effect until they persist. Thus cumulative currency movements that take place over a period of years will have an impact on trade flows that will persist even if the currency subsequently moves in a contrary direction for a temporary period.

Financial analysts need to consider a wide range of variables, both in the domestic economy and abroad, to assess relative GDP growth rates and then superimpose on those rates currency considerations to ascertain whether they reinforce other cyclical forces or counteract them. Generally, GDP growth differentials in global economic growth rates between countries have the most immediate and straightforward effect because domestic changes raise or reduce imports and foreign economic activity changes raise or reduce exports. Currency moves have a more complex and, despite the drama of short-term currency moves, a more gradual effect.

EXAMPLE 6-6 External Trade

1. Imports generally respond to:
   A. the level of exports.
   B. domestic industrial policy.
   C. the pace of domestic GDP growth.
2. Exports generally respond to:
   A. the level of unionization.
   B. the pace of global growth.
   C. the pace of domestic GDP growth.

Solution to 1: C is correct. As a part of aggregate demand, imports reflect the domestic needs for foreign goods, which vary together with domestic economic growth.

Solution to 2: B is correct. Exports reflect the foreign demands on domestic output, which depend on the conditions of global economy.

3. THEORIES OF THE BUSINESS CYCLE

Business cycles have been recognized since the early days of economic theory, and considerable effort has gone into identifying different cycles and explaining them. Until the 1930s,

however, the general view was that they were a natural feature of the economy and the pain of recessions was temporary. But the depth and severity of the 1930s downturn (known as the Great Depression) created a crisis in economic theory.

After the Great Depression (which began in 1929), the debate among various economic schools of thought (neoclassical, Austrian, and Keynesian) spurred important innovations in the way the business cycle was described and explained. Similarly, after the dramatic recessions triggered by the oil shocks of 1973 and 1979, the old paradigm was taken apart and new developments in economics and quantitative methods led to an improved understanding of short-term economic dynamics. In this section, we review and summarize some of the main theories.

3.1. Neoclassical and Austrian Schools

Neoclassical analysis relied on the concept of general equilibrium; that is, all markets will reach equilibrium because of the "invisible hand," or free market, and the price will be found for every good at which supply equals demand. All resources are used efficiently based on the principle of marginal cost equaling marginal revenue, and no involuntary unemployment of labor or capital takes place. In practice, because the neoclassical school provides that the invisible hand will reallocate capital and labor so that they will be used to produce whatever consumers want, it does not allow for fluctuations found in the aggregate economic activity. If a shock of any origin shifts either the aggregate demand curve or the aggregate supply curve, the economy will quickly readjust and reach its equilibrium via lower interest rates and lower wages.

Neoclassical economists rely on **Say's law**: All that is produced will be sold because supply creates its own demand. French economist J. B. Say pointed out that if something is produced, the capital and labor used for that production will have to be compensated. This compensation of the factors (interest for capital and wages for labor) creates purchasing power in the sense that the workers receive a paycheck and thus can buy goods and services they need. Widespread declines in demand would be strictly temporary.

In the neoclassical school, a massive crisis, such as the Great Depression of the 1930s with widespread unemployment of more than 20 percent throughout the industrialized world, is impossible. Yet it happened. The crisis started in the United States and successively affected many other countries.

The 1929 crisis helped introduce a breakthrough in economic theory because the crisis touched many sectors at the same time and in a dramatic fashion. Because the neoclassical theory denied the possibility of a prolonged depression, it could not be used to explain how to fight such a depression. The main adjustment mechanism proposed by the neoclassical school—cuts in wages—was difficult to achieve and, as we will see, was questioned by the Keynesian school.

The Austrian school, including Friedrich von Hayek and Ludwig von Mises, shared some views of the neoclassical school, but focused more on two topics that were largely unimportant in the neoclassical framework: money and government. Money was not necessary in the neoclassical model, because the exchange of goods and services could occur in the form of barter and still reach general equilibrium. Money was seen just as a way to simplify exchange. Similarly, the role of government in the neoclassical model was quite limited because the economy could take care of itself and little else was needed of the government besides upholding the law and securing the borders.

Von Hayek argued that fluctuations are caused by governments that try to increase GDP and employment (thus perhaps increasing voters' consensus) by adopting expansionary

monetary policies. Governments lower the market interest rate below its natural value (aggregate demand shifts right) and thus lead companies to overinvest (an inflationary gap). Once companies realize that they have accumulated too much equipment and too many structures, they will suddenly stop investing, which depresses aggregate demand (aggregate demand shifts left dramatically) and causes a crisis throughout the economy. To reach the new equilibrium, all prices, including wages, must decrease.

As a result of manipulating interest rates, the economy exhibits fluctuations that would not have happened otherwise. Therefore, Austrian economists advocate limited government intervention in the economy, lest the government cause a boom-and-bust cycle. The best thing to do in the recession phase is to allow the necessary market adjustment to take place as quickly as possible.

The Austrian school has a theory of what causes the business cycle: It is misguided government intervention. The neoclassical school does not have a theory of the business cycle, and the closest it gets to it is Joseph Schumpeter's creative destruction theory, which shows cycles within industries as a result of technological progress but no economy-wide fluctuations.[2] Schumpeter formulated a theory of innovations, which explained cycles limited to individual industries: When an inventor comes up with a new product (e.g., the digital music player in recent decades) or a new, better way to produce an existing good or service (e.g., radio frequency identification [RFID] tracking of inventories), then the entrepreneur that introduces the new discovery will likely have bigger profits and may drive the existing producers out of business. Therefore, innovations can generate crises that involve only the industry affected by the new invention. Neoclassical economics recognizes that business cycles exist but treats them as temporary disequilibria.

3.2. Keynesian and Monetarist Schools

The Keynesian and monetarist schools of economic thought have been among the most influential. Their prescriptions concerning the business cycle are discussed in the following sections.

3.2.1. Keynesian School

As previously mentioned, if a recession occurs, the neoclassical and Austrian schools argue in general that no government intervention is needed. Unemployment and excess supply of goods will be solved by allowing market prices to decrease (including wages) until all markets clear: Supply equals demand and factors of production are fully employed.

British economist John Maynard Keynes[3] disagreed with both neoclassical and Austrian views. He observed that a generalized price and wage reduction (solely brought about through market forces), necessary to bring markets back to equilibrium during a recession, would be hard to attain. For example, workers may not want to see their nominal compensation decrease, because nobody likes a pay cut.

But Keynes thought that even if workers agreed to accept lower salaries, this situation might exacerbate the crisis by reducing aggregate demand rather than solving it, because lower

[2]Joseph Alois Schumpeter was born in Austria and studied with members of the Austrian school, such as Menger and Hayek, but he was more neoclassical than Austrian in the economic sense. He taught in the United States for many years.

[3]John Maynard Keynes's name is often mentioned in full, with first and middle name, to avoid confusion with his father, John Neville Keynes, who was also an economist.

salary expectations would shift aggregate demand left. For example, if wages fell, workers would need to cut back on their spending. This response would cause a further contraction in the demand for all sorts of goods and services, starting with the more expensive items, such as durable goods, and moving in a domino effect through the economy (the downward spiral of the AD curve continuously shifting left, as mentioned earlier).

Further, Keynes believed there could be circumstances in which lower interest rates would not reignite growth because business confidence or "animal spirit" was too low. As a consequence, Keynes advocated government intervention in the form of fiscal policy. While he accepted the possibility that markets would reach the equilibrium envisioned by neoclassical and Austrian economists over the long run, he famously quipped that "in the long run, we are all dead"; that is, the human suffering is excessive while waiting for all shocks to be absorbed and for the economy to return to equilibrium.

When crises occur, the government should intervene to keep capital and labor employed by deliberately running a larger fiscal deficit. This intervention would limit the damages of major recessions. Although this concept continues to be a highly politically charged debate, many economists agree that government expenditure can limit the negative effect of major economic crises in the short term. Three practical criticisms are often expressed about Keynesian fiscal policy:

1. Fiscal deficits mean higher government debt that needs to be serviced and repaid eventually. There is a danger that government finances could move out of control.
2. Keynesian cyclical policies are focused on the short term. In the long run, the economy may come back and the presence of the expansionary policy may cause it to overheat—that is, to have unsustainably fast economic growth, which causes inflation and other problems. This result is because of the typical lags involved in expansionary policy taking effect on the economy.
3. Fiscal policy takes time to implement. Quite often, by the time stimulatory fiscal policy kicks in, the economy is already recovering. (Monetary policy determines the available quantities of money and loans in an economy.)

Keynes's writings did not advocate a continuous presence of the government in the economy, nor did he suggest using economic policy to fine-tune the business cycle. He only advocated decisive action in case of a serious economic crisis, such as the Great Depression.

The Perspective of Hyman Minsky

A different view of business cycles came from Hyman Minsky. His view had something in common with the Austrian school and something in common with Keynes. Minsky believed that excesses in financial markets exacerbate economic fluctuations. For example, a rapid growth of credit, often given to risky ventures in the late expansion phase of the cycle, will be followed by a credit crunch during the downswing phase. Minsky traced excesses to a type of complacency in which people underestimate the risk of events that have not occurred in a while. Therefore, if the economy has been in a long expansion, people may think that the market works very well and that the expansion

will last forever—that is, extrapolating past experiences. In this sense, Minsky could be seen as a precursor of behavioral finance, which is the branch of finance that studies how cognition biases, such as overconfidence and short memory, induce investors to be overconfident and make suboptimal choices.

The term **Minsky moment*** has been coined for a point in the business cycle when, after individuals become overextended in borrowing to finance speculative investments, people start realizing that something is likely to go wrong and a panic ensues leading to asset sell-offs. The subprime crisis that affected many industrialized countries starting in 2008 has been represented as a Minsky moment because it came after years in which risk premiums (e.g., the differentials, or spreads, between very risky bonds and very safe bonds) were at historically low levels. Typically, low risk premiums suggest that no adverse events are expected; in other words, investors believe that because the economy and the markets have been enjoying a protracted expansion, there is no reason to worry about the future. As a consequence, many market observers suggest that business cycles are being tamed. This kind of view of the world leads people to underestimate risk—for example, by not doing the appropriate diligent research before granting a loan or before purchasing a security—in a word, complacency.

The Minsky moment has been compared with a cartoon character who walks off a cliff without realizing that he is doing so. When he looks down and sees that he is walking on thin air, he panics and falls to the bottom of the canyon—just like the world economy in 2008.

A warning in 2005 by Alan Greenspan, then chairman of the U.S. Federal Reserve,† was not taken seriously by market participants. Greenspan said that historically, extended periods of low risk premiums have always ended badly. This warning was to a large extent in line with Minsky's view that people tend to extrapolate the recent past, and if little volatility has occurred recently, they may think that low volatility will persist indefinitely.

Some market analysts, after the crisis, pointed out that not even Greenspan knew how accurate his remarks would turn out to be. As explained by Reinhardt and Rogoff (2009) and Siegel (2010), when rates were very low, investors tried to get extra yield by investing in complex and potentially more risky assets, such as securitized subprime mortgages and collateralized debt obligations whose creditworthiness was much lower than expected.

As soon as the economy started having difficulties, the value of many risky securities dropped dramatically, placing investors globally into a state of panic and causing a dramatic fall in aggregate demand. The crisis is another example of why thinking that a new era has started and that the things learned from analyzing the past no longer apply is generally a very costly mistake.

*Paul McCulley (for example, see McCulley 2010) originated this expression.

†Alan Greenspan, chairman of the U.S. Federal Reserve at the time, remarked in 2005: "Thus, this vast increase in the market value of asset claims is in part the indirect result of investors accepting lower compensation for risk. Such an increase in market value is too often viewed by market participants as structural and permanent. To some extent, those higher values may be reflecting the increased flexibility and resilience of our economy. But what they perceive as newly abundant liquidity can readily disappear. Any onset of increased investor caution elevates risk premiums and, as a consequence, lowers asset values and promotes the liquidation of the debt that supported higher asset prices. This is the reason that history has not dealt kindly with the aftermath of protracted periods of low risk premiums."

3.2.2. Monetarist School

The monetarist school, generally identified with Milton Friedman, objected to Keynesian intervention for four main reasons:

1. The Keynesian model does not recognize the supreme importance of the money supply. If the money supply grows too fast, there will be an unsustainable boom, and if it grows too slowly, there will be a recession. Friedman focused mainly on broad measures of money, such as M2.
2. The Keynesian model lacks a complete representation of utility-maximizing agents and is thus not logically sound.
3. Keynes's short-term view fails to consider the long-term costs of government intervention (e.g., growing government debt and high cost of interest on this debt).
4. The timing of governments' economic policy responses is uncertain, and the stimulative effects of a fiscal expansion may take effect after the crisis is over, and thus cause more harm than good.[4]

Therefore, monetarists advocate a focus on maintaining steady growth of the money supply, and otherwise a very limited role for government in the economy. Fiscal and monetary policy should be clear and consistent over time, so all economic agents can forecast government actions. In this way, the uncertainty of economic fluctuations would not be increased by any uncertainty about the timing and magnitude of economic policies and their lagged effects.

According to the monetarist school, business cycles may occur both because of exogenous shocks and because of government intervention. It is better to let aggregate demand and supply find their own equilibrium than to risk causing further economic fluctuations. However, a key part of monetarist thought is that the money supply needs to continue to grow at a moderate rate. If it falls, as occurred in the 1930s, the economic downturn could be severe, whereas if money grows too fast, inflation will follow.

3.3. The New Classical School

Starting in the 1970s, economists such as Robert Lucas started questioning the foundations of the models used to explain business cycles. Among other things, Lucas agreed with Friedman (1968) and pointed out that the models should try to represent the actions of economic agents with a utility function and a budget constraint, just like the models used in microeconomics. This approach has come to be known as **new classical macroeconomics**—an approach to macroeconomics that seeks the macroeconomic conclusions of individuals maximizing utility on the basis of rational expectations and companies maximizing profits. The assumption is made that all agents are roughly alike, and thus solving the problem of one agent is the same as solving the problems of millions of similar agents (or the per capita income and consumption of the average agent).

The new classical models are dynamic in the sense of describing fluctuations over many periods and present general equilibrium in the sense of determining all prices rather than one price. The models by Edward C. Prescott and Finn E. Kydland, who are among the pioneers of this approach, have an economic agent that has to face external shocks (e.g., as a result of

[4]Markets may react differently to changes in interest rates and other tools of monetary policy. There is a long chain of events from the time when interest rates are cut, to when banks change the rates they charge clients, to when a company sees that rates are lower and thus decides to invest in new equipment, to when the equipment is finally purchased. Therefore, by the time these events all happen, the economy may be in expansion and the new investment may lead the economy to overheating.

changes in technology, tastes, or world prices) and thus optimizes its choices to reach the highest utility. If all agents act in similar fashion, the markets will gradually adjust toward equilibrium.

3.3.1. Models without Money: Real Business Cycle Theory

New classical economists comment that some policy recommendations made in the past were rather illogical: for example, if everybody knows that in a recession the government will give out low-rate loans to corporations that want to invest in new equipment and structures, why would any reasonable company invest outside recessions unless absolutely required to? Obviously, if most companies thought that, they would stop investing, thus causing a recession that otherwise would not have occurred. Essentially, the government's anticyclical policy could cause a recession.

Because, just like the neoclassical models, the initial new classical models did not include money, they were called real business cycle models (often abbreviated as RBC models). Cycles have real causes, such as the aforementioned changes in technology, whereas monetary variables, such as inflation, are assumed to have no effect on GDP and unemployment.[5]

RBC models of the business cycle conclude that expansions and contractions represent efficient operation of the economy in response to external real shocks. Because the level of economic activity at any time is consistent with maximizing expected utility, the policy recommendation of RBC theory is for government *not* to intervene in the economy with discretionary fiscal and monetary policy.

Critics of RBC models often focus on the labor market. Because RBC models rely on efficient markets, it follows that unemployment can only be short term: apart from frictional unemployment,[6] if markets are efficient, a person who does not have a job can only be a person who does not want to work. If people are unemployed, in the context of efficient markets, they just need to lower their wage rate until they find an employer who hires them. This assumption is logical because if markets are perfectly flexible, all markets must find equilibrium and full employment.

Therefore, as suggested particularly by the earliest RBC models, a person is unemployed because he or she is asking for wages that are too high, or in other words, this person's utility function is maximized by having more leisure (e.g., free time to visit museums, watch games on TV, and enjoy time with friends) and less consumption (which could be increased by giving up some leisure and finding a job). However, the observation that during a recession many people are eagerly searching for jobs and are unable to find employment despite dropping their asking wages substantially suggests that this theory is unrealistic.

Although many find this explanation unconvincing, RBC theorists argue that, undeniably, markets would clear if people were rational and avoided unrealistic expectations of earnings or simply enjoyed their leisure accompanied by optimally meager consumption.

An interesting feature of RBC models is that they give aggregate supply a more prominent role than many other theories. For example, supply has a limited importance in the Keynesian

[5]See Plosser (1989) and Romer (2005, chap. 4) for an introduction to RBC models. Basically, RBC models assume that economic agents are fully rational and that markets function with no imperfection or friction. As a consequence, any changes in monetary aggregates or other monetary policies will promptly cause changes in price levels and other variables without affecting real GDP or employment.

[6]Frictional unemployment indicates the people who in economic statistics appear to be unemployed, but in reality are just moving between jobs. Because there is always someone who has just left a job and is about to start a new one, or someone who just entered the job market and has already found a job but has not started yet, we know that a small part of the statistically unemployed are people between jobs.

theory, probably because Keynes was more concerned with the Great Depression, which was largely a crisis of aggregate demand. RBC models show that supply shocks, such as advances in technology or changes in the relative prices of inputs, cause the aggregate supply (AS) curve to shift left. A new technology can change potential GDP, for example, thus moving long-run AS to the right. Adjustment will be needed because not all companies can adopt the new technology at once, and therefore short-run AS will not jump to the new equilibrium immediately. Similarly, an increase of energy prices shifts short-run AS to the left (higher prices and lower GDP). In the long run, companies and households can learn to use less of the expensive energy inputs (substitution effect), and therefore long-run AS will shift right (higher GDP) if the economy learns to produce more goods with less energy.

3.3.2. Models with Money

Inflation is often seen as a cause of business cycles, because when monetary policy ends up being too expansionary, the economy grows at an unsustainable pace—creating an inflationary gap. The result is that, for example, suppliers cannot keep up with demand. In this environment, prices will tend to grow faster than normal—that is, inflation.

As a consequence, the central bank will often intervene to limit inflation by tightening monetary policy, which generally means increasing interest rates, so that the cost of borrowing will be higher and demand for goods and services will slow down (a leftward shift in aggregate demand caused by the higher cost of money). This response will decrease equilibrium GDP and can result in a recession.

Given that inflation appears to trigger policy responses from central banks, it is an important part of modern business cycles. Therefore, it can be helpful to use models that include money to explain economic growth. As mentioned earlier, RBC models assume that transactions could occur with barter, and thus do not explicitly include money. More recent dynamic general equilibrium models (for example, Christiano, Eichenbaum, and Evans, 2005) include money and inflation.

Monetary policy can be incorporated into dynamic general equilibrium models with money. In one type of model, the economy receives shocks from changes in technology and consumer preferences (like in the RBC case), but can also receive shocks from monetary policy, which sometimes can tame the business cycle and at other times may exacerbate it.

Another group of dynamic general equilibrium models are the **neo-Keynesians** or **new Keynesians**.[7] Like the new classical school, the neo-Keynesian school attempts to place macroeconomics on sound microeconomic foundations. In contrast to the new classical school, the neo-Keynesian school assumes slow-to-adjust (sticky) prices and wages. The neo-Keynesian models show that markets do not reach equilibrium immediately and seamlessly, but even small imperfections may cause markets to be in disequilibrium for a long time. As a consequence, government intervention as advocated in the 1930s by Keynes may be useful to eliminate unemployment and bring markets toward equilibrium.

The typical example of these imperfections, which also appeared in Keynes's work, is that workers do not want their wages to decrease to help the market reach a new equilibrium (i.e., wages are often downwardly sticky).[8] Another possibility that some economists suggested in the 1980s is called the "menu costs" explanation: It is costly for companies to continuously

[7]For an introduction to neo-Keynesian models, see Romer (2005, chap. 5) and Mankiw (1989).

[8]As mentioned earlier, Keynes thought that even if workers agreed to accept lower wages, this might exacerbate the crisis rather than solving it because lower wage expectations would shift AD left.

adjust prices to make markets clear, just like it would be costly for a restaurant to print new menus daily with updated prices.[9] Another explanation is that every time an economic shock hits a company, the company will need some time to reorganize its production.

EXAMPLE 6-7 Real Business Cycle Models

1. The main difference between new classical (RBC) and neo-Keynesian models is that the new classical models:
   A. are monetarist.
   B. use utility-maximizing agents, whereas neo-Keynesian models do not.
   C. assume that prices adjust quickly to changes in supply and demand, whereas neo-Keynesians assume that prices adjust slowly.
2. Basic RBC models focus on the choices of a typical individual, who can choose between consuming more (thus giving up leisure) and enjoying leisure more (thus giving up consumption). What causes persistent unemployment in this model?
   A. Contractionary monetary policy causes a shock to real variables.
   B. The economy returns to equilibrium promptly; thus persistent unemployment does not exist.
   C. The utility function: Individuals who prefer leisure much more than consumption will forgo consumption and instead choose unemployment to enjoy more leisure when the market salary is low.

Solution to 1: C is correct. A key feature of Keynesian macroeconomics is the stickiness of prices. In contrast, classical views assume flexible price adjustments that ensure market clearing. Modern theories always assume rational economic representative agents in the economy as the microeconomic foundation of macroeconomics. Thus they are the "new" models.

Solution to 2: C is correct. Shocks in the standard new classical model can have only a temporary effect; thus A is not the right answer. Unemployment can still exist when the labor market is cleared, so a rational explanation is provided in C.

In recent years, a consensus concerning business cycles has gradually started building in macroeconomics. It is too early to say that economists agree on all causes of and remedies for business fluctuations, but at least an analytical framework has emerged that encompasses both new classical and neo-Keynesian approaches. Woodford (2009), among others, shows that new research seems to be leading to a unified approach.

The debate about business cycles often receives a politically partisan treatment in the press because some people are generally against government intervention in the economy (for

[9]Clearly, both this example and the "menu costs" name were initially envisioned before personal computers and laser printers became affordable and widely used. Still, one can imagine the cost for a store owner to replace the price tags on every item in the store on a daily basis, and also how this would confuse shoppers.

example, because it may lead to large deficits) and others are in favor (for example, because it may alleviate the effects of a large economic shock). It is important to base investment decisions on analysis and not on politics; the financial analyst must try as much as possible to set personal biases aside.

However, there is little doubt that central banks are very actively trying to manage the business cycle by raising interest rates when the economy becomes too hot and inflation accelerates and cutting rates when the economy is weak. In the 2008–2009 downturn, when official interest rates approached zero, this policy was extended to include so-called quantitative easing to try to lower interest rates further out on the yield curve to stimulate the economy.

Analyzing Government Expenditure

Simple criteria for the financial analyst wondering if a government's expenditure is excessive (i.e., unsustainably high or of an inappropriate composition) include the following:*

1. Does the government always have a deficit no matter the cyclical phase, or does it have surpluses during economic booms?
2. Does the government have a deficit because of a defined series of necessary investments that will improve the productivity of the country, or is it spending most of its money in salaries for patronage employees and on infrastructure of questionable uses?
3. Is the growth rate of debt (government budget deficit as a percentage of GDP) higher than GDP growth? If so, the debt level will not likely be sustainable.

When government expenditures are excessive, inflation often follows. After that, a recession may occur because the central bank takes necessary measures to slow down an overheated economy. That is, if government purchases increase aggregate demand too much, thus causing inflation (expansionary fiscal policy), the central bank will intervene to stop prices from increasing too quickly (tightening or contractionary monetary policy).

*For a more formal and data-rich approach, see Reinhart and Rogoff (2009).

4. UNEMPLOYMENT AND INFLATION

Many governments state economic policy objectives related to limiting the rate at which citizens are unemployed and containing price inflation (i.e., preserving the purchasing power of a domestic currency). The relationships of these variables to the business cycle are discussed in the following sections. In general, unemployment is at its highest just as the recovery starts and is at its lowest at the peak of the economy.

4.1. Unemployment

A typical cause of business cycle downturns is a tight labor market—that is, one with low unemployment. An overheated economy leads to inflation when unemployment is very low. Workers ask for higher wages because they expect prices of goods and services to keep going up, and at the same time they have market power against employers because there are few available workers to be hired. This upward pressure on wages coupled with the impact of wage escalator clauses (automatic increases in wages as the consumer price index grows) triggers a price-wage inflationary spiral. This issue was a particular problem in industrialized countries in the 1960s and 1970s and remains an issue today.

A key aspect in this process is inflation expectations. Because inflation expectations are high, the request for higher wages is stronger, which induces employers to increase prices in advance to keep their profit margins stable. This avalanche process grows with time. Central banks act, sometimes drastically, to slow down the economy and reset inflationary expectations throughout the economy at a low level so that if everyone expects low inflation, the inflationary spiral itself will stop. An effect of these policies can be a deep recession. Therefore, whenever a financial analyst sees signs of a price-wage spiral in the making, a reasonable response would be to consider the effect of both high inflation and sharp tightening of monetary policy.

This example shows that measures of labor market conditions are important in assessing whether an economy is at risk of a cyclical downturn.

The following are the definitions of a few terms that are used to summarize the state of the labor market:

- **Employed**: number of people with a job. This figure normally does not include people working in the informal sector (e.g., unlicensed cab drivers, illegal workers, etc.).
- **Labor force**: number of people who either have a job or are actively looking for a job. This number excludes retirees, children, stay-at-home parents, full-time students, and other categories of people who are neither employed nor actively seeking employment.
- **Unemployed**: people who are actively seeking employment but are currently without a job. Some special subcategories include:
  - Long-term unemployed: people who have been out of work for a long time (more than three to four months in many countries) but are still looking for a job.
  - Frictionally unemployed: people who are not working at the time of filling out the statistical survey because they just left one job and are about to start another job. That is, the frictionally unemployed have a job waiting for them and are not 100 percent unemployed; it is just that they have not started the new job yet.
- **Unemployment rate**: ratio of unemployed to labor force.
- **Activity ratio** (or participation ratio): ratio of labor force to total population of working age (i.e., those between 16 and 64 years of age).
- **Underemployed**: person who has a job but has the qualifications to work a significantly higher-paying job. For example, a lawyer who takes a job in a bookstore could call herself underemployed. This lawyer would count as employed for the computation of the unemployment rate (she does have a job, even if it may not be her highest-paying job). Although the unemployment rate statistic is criticized for not taking the issue of underemployment into account, it may be difficult to classify whether a person is truly

underemployed; for example, the lawyer may find legal work too stressful and prefers working at the bookstore. However, data for part-time working is sometimes a good proxy.

- **Discouraged worker**: person who has stopped looking for a job. Perhaps because of a weak economy, the discouraged worker has given up seeking employment. Discouraged workers are statistically outside the labor force (similar to children and retirees), which means they are not counted in the official unemployment rate. During bad recessions, the unemployment rate may actually decrease because many discouraged workers stop seeking work. It is important to observe the participation rate together with the unemployment rate to understand if unemployment is decreasing because of an improved economy or because of an increase in discouraged workers. Discouraged workers and underemployed people may be considered examples of hidden unemployment.
- **Voluntarily unemployed**: person voluntarily outside the labor force, such as a jobless worker refusing an available vacancy for which the wage is lower than the individual's threshold or someone who retired early.

4.1.1. The Unemployment Rate

The unemployment rate is certainly the most quoted measure of unemployment; it attempts to measure those people who have no work but would work if they could find a job, generally stated as a percentage of the overall workforce. In the United States, the indicator emerges from a monthly survey of households by the U.S. Bureau of Labor Statistics, which asks how many household members have jobs and how many of working age do not but are seeking work. Other statistical bureaus rely on other sources for the calculation, using claims for unemployment assistance, for instance, or their equivalent. Some measure the workforce simply as those of working age, regardless of whether they are ready or willing to work. These differences can make precise international comparisons problematic. One solution is to use the International Labour Organization (ILO) statistics that try to estimate on a consistent basis. As indicated earlier, some statistical agencies add perspective with other measures—for example, what proportion of those who have ceased work are discouraged, underemployed, or have opted out of the workforce for other reasons or what proportion are working part-time.

Although these various unemployment measures provide insight into the state of the economy, they are inaccurate in pointing to cyclical directions on at least two counts, both of which make unemployment a lagging economic indicator of the business cycle.

In the first place, the unemployment rate tends to point to a past economic condition—that is, it lags the cycle—because the labor force expands and declines in response to the economic environment. Compounding the inaccuracy, when times get hard, discouraged workers cease searching for work, reducing the number typically counted as unemployed and making the jobs market look stronger than it really is. Conversely, when the jobs market picks up, these people return to the search, and because they seldom find work immediately, they at least initially raise the calculation of those unemployed, giving the false impression of the lack of recovery in the jobs market, when, in fact, it is the improvement that brought these people back into the job search in the first place. Sometimes this cyclical flow of new job seekers is so great that the unemployment rate actually rises even as the economic recovery gains momentum. Those agencies that measure the workforce in terms of the working-age population avoid this bias, because this demographic (working-age population) remains more or less

constant regardless of the state of the labor market. But this approach introduces biases of its own, such as counting as unemployed people who have severe disabilities and could never seek work.

The second reason the unemployment indicator tends to lag the cycle comes from the typical reluctance of businesses to lay off people. The reluctance may stem from a desire to retain good workers for the long run, or just reflect constraints written into labor contracts that make layoffs expensive. The reluctance makes the various measures of unemployment rise more slowly as the economy slides into recession than they otherwise might. Then as the recovery develops, a business waits to hire until it has fully employed the workers it has kept on the payroll during the recession; this delay causes decreases in the unemployment rate to lag in the cyclical recovery, sometimes for a long time.

4.1.2. Overall Payroll Employment and Productivity Indicators

To get a better picture of the employment cycle, practitioners often rely on more straightforward measures of payroll growth. By measuring the size of payrolls, practitioners sidestep such issues as the ebb and flow of discouraged workers and can more directly point to cyclical directions. These statistics, however, do have biases of their own. It is hard, for instance, to count employment in smaller businesses, which, especially in many developed economies, are the main drivers of employment growth. Still, there is a clear indication of economic trouble when payrolls shrink and a clear indication of recovery when they rise.

The application of other variables can also assist in understanding the employment situation and its use in determining cyclical directions. Two of the most straightforward are measures of hours worked, especially overtime, and the use of temporary workers. A business does not want to make mistakes with full-time staff, either hiring or firing. Thus, at the first signs of economic weakness, managers cut back hours, especially overtime. Such movements can simply reflect minor month-to-month production shifts, but if followed by cutbacks in part-time and temporary staff, the picture gives a strong signal of economic weakness, especially if confirmed by other independent indicators. Similarly, on the cyclical upswing, a business turns first to increases in overtime and hours. If a business then increases temporary staffing, it gives a good signal of economic recovery long before any movement in rehiring full-time staff again, especially if confirmed by independent cyclical indicators.

Productivity measures also offer insight into this cyclical process. Because productivity is usually measured by dividing output by hours worked, a business's tendency to keep workers on the payroll even as output falls usually prompts a reduction in measured productivity. If measures are available promptly enough, this sign of cyclical weakness might precede even the change in hours. This drop in productivity precedes any change in full-time payrolls. Productivity also responds promptly when business conditions improve and the business first begins to utilize its underemployed workers, which occurs earlier than any upturn in full-time payrolls.

On a more fundamental level, productivity can also pick up in response to technological breakthroughs or improved training techniques. As already mentioned, such changes affect potential GDP. If strong enough, they can negatively affect employment trends, keeping them slower than they would be otherwise by relieving the need for additional staff to increase production. But these influences usually unfold over decades and mean little to cyclical considerations, which unfold over years at most. What is more, there are few statistical indicators to gauge the onset of technological change, restraining analysts to the use of anecdotal evidence or occasional longitudinal studies.

EXAMPLE 6-8 Analyzing Unemployment

1. Comparisons of unemployment among countries:
   A. are impossible.
   B. show which countries are more prosperous.
   C. must take into account different unemployment measurement methods.
2. Unemployment frequently lags the business cycle because:
   A. it takes time to compile the employment data.
   B. businesses are reluctant to dismiss and hire workers.
   C. workers must give notice to employers before quitting jobs.
3. Productivity offers perspective on the business cycle by:
   A. showing the need for new employees.
   B. assessing the skill set of existing employees.
   C. measuring the intensity of work flow for existing employees.

Solution to 1: C is correct. Different countries use different statistical scope and ratio definitions, and these differences have to be reconciled before meaningful conclusions can be made from cross-country comparisons.

Solution to 2: B is correct. Besides labor hoarding by employers because of the costs related to hiring and firing, the variations of labor force over business cycles also contribute to this feature of the unemployment rate.

Solution to 3: C is correct. Because employers would like to keep the workforce relatively stable, productivity falls as output declines in a downturn because it is measured as the ratio of output over hours worked. Similarly, productivity rises as output recovers.

4.2. Inflation

The overall price level changes at varying rates during different phases of a business cycle. Thus, when studying business cycles, it is important to understand the statistics related to this phenomenon. In general, the inflation rate is procyclical (that is, it goes up and down *with* the cycle), but with a lag of a year or more.

Inflation refers to a sustained rise in the overall level of prices in an economy. Economists use various price indexes to measure the overall price level, also called the aggregate price level. The **inflation rate** is the percentage change in a price index—that is, the speed of overall price level movements. Investors follow the inflation rate closely, not only because it can help to infer the state of the economy but also because an unexpected change may result in a change in monetary policy, which can in turn have a large and immediate impact on asset prices. In developing countries, very high inflation rates can lead to social unrest or even shifts of political power, which constitutes political risk for investments in those economies.

Central banks, the monetary authority in most modern economies, monitor the domestic inflation rates closely when conducting monetary policy, which in turn determines the available quantities of money and loans in an economy. A high inflation rate combined with fast economic growth and low unemployment usually indicates the economy is overheating,

which may trigger some policy movements to cool it down. However, if a high inflation rate is combined with a high level of unemployment and a slowdown of the economy—an economic state known as **stagflation** (for stagnation plus inflation)—the economy will typically be left to correct itself, because no short-term economic policy is thought to be effective.

4.2.1. Deflation, Hyperinflation, and Disinflation

There are various terms related to the levels and changes of the inflation rate.

- **Deflation**: a sustained decrease in aggregate price level, which corresponds to a negative inflation rate—that is, an inflation rate of less than 0 percent.
- **Hyperinflation**: an extremely fast increase in aggregate price level, which corresponds to an extremely high inflation rate—for example, 500 to 1,000 percent per year.
- **Disinflation**: a decline in the inflation rate, such as from around 15 to 20 percent to 5 or 6 percent. Disinflation is very different from deflation because even after a period of disinflation, the inflation rate remains positive and the aggregate price level keeps rising (although at a slower speed).

Inflation means that the same amount of money can purchase less real goods or services in the future. So, the value of money or the purchasing power of money decreases in an inflationary environment. When deflation occurs, the value of money actually increases. Because most debt contracts are written in fixed monetary amounts, the liability of a borrower also rises in real terms during deflation. As price levels fall, the revenue of a typical company also falls during a recession. Facing increasing real debt, a company that is short of cash usually cuts its spending, investment, and workforce sharply. Less spending and high unemployment then further exacerbate the economic contraction. To avoid getting too close to deflation, the consensus on the preferred inflation rate is around 2 percent per year for developed economies. Deflation occurred in the United States during the Great Depression and briefly during the recession of 2008–2009 following the global financial crisis. Since the late 1990s, Japan has experienced several episodes of deflation.

Hyperinflation usually occurs when large-scale government spending is not backed by real tax revenue and the monetary authority accommodates government spending with unlimited money supply. Hyperinflation is often triggered by the shortage of supply created during or after a war, economic regime transition, or prolonged economic distress of an economy caused by political instability. During hyperinflation, people are eager to change their cash into real goods because prices are rising very fast. As a result, money changes hands with extremely high frequency. The government also has to print more money to support its increased spending. As more cash chases a limited supply of goods and services, the rate of price increases accelerates. After World War I, a famous case of hyperinflation occurred in Germany from 1923 to 1924. During the peak of this episode, prices doubled every 3.7 days. After World War II, Hungary experienced a severe hyperinflation during which prices doubled every 15.6 hours at its peak in 1946. In 1993, the inflation rate in Ukraine peaked at 10,155 percent per year. In January 1994, the *monthly* inflation rate peaked at 313 million percent in Yugoslavia. The most recent hyperinflation in Zimbabwe reached a peak of *monthly* inflation at 79.6 billion percent in the middle of November 2008. Because the basic cause for hyperinflation is too much money in circulation, regaining control of the money supply is the key to ending hyperinflation.

Exhibit 6-4 shows two recent episodes of disinflation in selected countries around the world. The first episode happened during the early 1980s. Because of the two oil crises in

EXHIBIT 6-4 Two Episodes of Disinflation around the World: Annual Inflation Rates

	First Episode					Second Episode			
Year	1979	1980	1983	1984	1985	1990	1991	1998	1999
Country									
Australia	9.1	10.2	10.1	3.9	6.7	7.3	3.2	0.9	1.5
Canada	9.1	10.1	5.9	4.3	4.0	4.8	5.6	1.0	1.7
Finland	7.5	11.6	8.4	7.1	5.2	6.1	4.3	1.4	1.2
France	10.6	13.6	9.5	7.7	5.8	3.2	3.2	0.6	0.5
Germany	4.0	5.4	3.3	2.4	2.1	2.7	4.0	1.0	0.6
Italy	14.8	21.1	14.6	10.8	9.2	6.5	6.3	2.0	1.7
Japan	3.7	7.8	1.9	2.3	2.0	3.1	3.3	0.7	–0.3
South Korea	18.3	28.7	3.4	2.3	2.5	8.6	9.3	7.5	0.8
Spain	15.7	15.6	12.2	11.3	8.8	6.7	5.9	1.8	2.3
Sweden	7.2	13.7	8.9	8.0	7.4	10.4	9.4	–0.3	0.5
United Kingdom	13.4	18.0	4.6	5.0	6.1	7.0	7.5	1.6	1.3
United States	11.3	13.5	3.2	4.3	3.5	5.4	4.2	1.6	2.2
Average	10.4	14.1	7.2	5.8	5.3	6.0	5.5	1.6	1.2
G-7 countries	9.6	12.5	4.6	4.6	3.9	4.8	4.4	1.3	1.4

Source: Organization for Economic Cooperation and Development (OECD).

the 1970s, many countries around the world were experiencing high levels of inflation. In Exhibit 6-4, the annual inflation rates in most counties around 1980 ranged between 10 and 20 percent. Even though this level is still far from hyperinflation, it generated social pressure against inflationary monetary policy. At the cost of a severe recession early in the 1980s, these countries brought inflation rates down to around 5 percent on average by 1985. In the first years of the 1990s, inflationary experience varied widely in world markets as some countries, such as the United States and the United Kingdom, entered recessions while others boomed. However, from the beginning to the end of the decade, there was a broad-based decline in inflation rates; in some countries annual inflation rates were below 2 percent by the end of the decade. In many countries, the decline in inflation was attributed to high productivity growth rates.

4.2.2. Measuring Inflation: The Construction of Price Indexes

Because the inflation rate is measured as the percentage change of a price index, it is important to understand how a price index is constructed so that the inflation rate derived from that index can be accurately interpreted. A **price index** represents the average prices of a basket of goods and services, and various methods can be used to average the different prices. Exhibit 6-5 shows a simple example of the change of a consumption basket over time.

EXHIBIT 6-5 Consumption Basket and Prices over Two Months

Time	January 2010		February 2010	
Goods	Quantity	Price	Quantity	Price
Rice	50 kg	¥3/kg	70 kg	¥4/kg
Gasoline	70 liters	¥4.4/liter	60 liters	¥4.5/liter

For January 2010, the total value of the consumption basket is:

$$\text{Value of rice} + \text{Value of gasoline} = (50 \times 3) + (70 \times 4.4) = ¥458$$

A price index uses the relative weight of a good in a basket to weight the price in the index. Therefore, the same consumption basket in February 2010 is worth:

$$\text{Value of rice} + \text{Value of gasoline} = (50 \times 4) + (70 \times 4.5) = ¥515$$

The price index in the base period is usually set to 100. So if the price index in January 2010 is 100, then the price index in February 2010 is:

$$\text{Price index in February 2010} = \frac{515}{458} \times 100 = ¥112.45$$

$$\text{The inflation rate} = \frac{112.45}{100} - 1 = 0.1245 = 12.45\%$$

A price index created by holding the composition of the consumption basket constant is called a **Laspeyres index**. Most price indexes around the world are Laspeyres indexes because the survey data on the consumption basket are available only with a lag. In many countries, the basket is updated every five years. Because most price indexes are created to measure the cost of living, simply using a fixed basket of goods and services has three serious biases:

1. *Substitution bias.* As the price of one good or service rises, people may replace it with another good or service that has a lower price. This substitution will result in an upward bias in the measured inflation rate based on a Laspeyres index.
2. *Quality bias.* As the quality of the same product improves over time, it better satisfies people's needs and wants. One such example is the quality of cars. Over the years, the prices of cars have been rising but the safety and reliability of cars have also been enhanced. If not adjusted for quality, the measured inflation rate will experience another upward bias.
3. *New product bias.* New products are frequently introduced that are not included in a fixed basket of goods and services. In general, this situation again creates an upward bias in the inflation rate.

It is relatively easy to resolve the quality bias and the new product bias. Many countries adjust for the quality of the products in a basket, a practice called hedonic pricing. New products can be introduced into the basket over time. The substitution bias can be somewhat resolved by using chained price index formula. One such example is the **Fisher index**, which is the geometric mean of the Laspeyres index and the **Paasche index**. The latter is an index formula using the current composition of the basket. Using the consumption basket for February 2010 in Exhibit 6-5, the value of the Paasche index is:

$$\text{Paasche index}_{02/2010} = I_P = \frac{(70 \times 4) + (60 \times 4.5)}{(70 \times 3) + (60 \times 4.4)} \times 100$$
$$= \frac{550}{474} \times 100 = 116.03$$

The value of the Fisher index is:

$$\text{Fisher index}_{02/2010} = \sqrt{I_P \times I_L} = \sqrt{116.03 \times 112.45} = 114.23$$

where I_L is the Laspeyres index.

4.2.3. Price Indexes and Their Usage

Most countries use their own consumer price index (CPI) to track inflation in the domestic economy. Exhibit 6-6 shows the different weights for various categories of goods and services in the consumer price indexes of different countries.

EXHIBIT 6-6 Consumption Baskets of Different Consumer Price Indexes

Country	Japan	China	India	Germany	U.S.	U.S.
Name of Index	CPI	CPI	CPI (UNME)	HICP	PCE	CPI-U
Year[a]	2005	2005	1984–1985	2008	2009	2007–2008
Category (%):						
Food and beverages	25.9	34	47.1	16.7	7.8	14.8
Housing and utilities	27.2	13	21.9	23.1	18.8	37.4
Furniture	3.4	6	2.0	6.1	2.5	4.6
Apparel	4.6	9	7.0	5.3	3.2	3.7
Medical care	4.5	10	2.5	4.4	16.2	6.5
Transportation and communication	13.9	10	5.2	17.6	9.1	20.1
Education and recreation	14.6	14	6.8	13.3	7.0[b]	9.5
Others	5.9	4	7.5	13.5	35.4	3.5

[a]The base year of the weights where it is appropriate.
[b]Recreation only.
Sources: Government websites and authors' calculations.

As shown in Exhibit 6-6, in different countries the consumer price indexes have different names and different weights on various categories of goods and services. For example, food weights are higher in the CPIs for China and India, but less for the developed countries. The scope of the index is also different among countries. For China, Japan, and Germany, the surveys used to collect data for CPI cover both urban and rural areas. The CPI for the United States covers only urban areas using a household survey, which is why it is called the CPI-U. In contrast, the **personal consumption expenditures** (PCE) price index covers all personal consumption in the United States using business surveys. The **producer price index** (PPI) is another important inflation measure. The PPI reflects the price changes experienced by domestic producers in a country. Because price increases may eventually pass through to consumers, the PPI can influence the future CPI. The items in the PPI include fuels, farm products (such as grains and meat), machinery and equipment, chemical products (such as drugs and paints), transportation equipment, metals, pulp and paper, and so on. These products are usually further grouped by stage-of-processing categories: crude materials, intermediate materials, and finished goods. Similar to the CPI, scopes and weights vary among countries. The differences in the weights can be much more dramatic for the PPI than for the CPI because different countries may specialize in different industries. In some countries, the PPI is called the **wholesale price index** (WPI).

As an important inflation indicator, many economic activities are indexed to a certain price index. For example, in the United States, a **Treasury Inflation-Protected Security** (TIPS) adjusts the bond's principal according to the U.S. CPI-U index. The terms of labor contracts and commercial real estate leases may adjust periodically according to the CPI. Recurring payments in business contracts can be linked to the PPI or its subindexes for a particular category of products.

Central banks usually use a consumer price index to monitor inflation. For example, the European Central Bank (ECB), the central bank for the European Union (EU), focuses on the harmonized index of consumer prices (HICP). Each member country in the EU first reports its own individual HICP, and then Eurostat, the statistical office for the EU, aggregates the country-level HICPs with country weights. There are exceptions, however. The Reserve Bank of India follows the inflation in India using the WPI. Because food items represent only about 27 percent in the India WPI (much lower than the 70 percent in the India rural CPI), the rural CPIs can rise faster than the WPI when there is high food price inflation. Besides the weight differences, the wholesale prices in the WPI also understate market prices because they do not take into account retail margins (markups). The choice of inflation indicator may also change over time. The central bank of the United States, known as the Federal Reserve (the Fed) once focused on the CPI-U produced by the Bureau of Labor Statistics under the U.S. Department of Labor. Because the CPI-U is a Laspeyres index and it has the previously discussed upward biases, the Fed switched in 2000 to the PCE index, a Fisher index produced by the Bureau of Economic Analysis under the U.S. Department of Commerce. The PCE index also has the advantage that it covers the complete range of consumer spending rather than just a basket.

Besides tracking inflation, financial analysts also use the price index to deflate GDP (i.e., to eliminate the price effect in nominal GDP data so as to identify trends in real economic growth). Many countries publish a particular price index, called the GDP deflator, for that purpose. Subindexes are also commonly available and may prove more valuable to an analyst with an interest in a particular industry or company.

Headline and Core Inflation

Headline inflation refers to the inflation rate calculated based on the price index that includes all goods and services in an economy. **Core inflation** usually refers to the inflation rate calculated based on a price index of goods and services except food and energy. Policy makers often choose to focus on the core inflation rate when reading the trend in the economy and making economic policies. The reason is that policy makers are trying to avoid overreaction to short-term fluctuations in food and energy prices that may not have a significant impact on future headline inflation.

The ultimate goal for policy makers is to control headline inflation, which reflects the actual cost of living. The fluctuations in the prices of food and energy are often the result of short-term changes in supply and demand. These changes in the prices of energy, particularly oil, are internationally determined and not necessarily reflective of the domestic business cycle. These imbalances may not persist, or even if some changes are permanent, the economy may be able to absorb them over time. These possibilities make headline inflation a noisy predictor. The core inflation rate may be a better signal of the trend in domestically driven inflation. To the extent that some trends in the headline inflation rate are permanent, policy makers need to pay attention to these as well.

Subindexes and Relative Prices

As mentioned previously, a subindex refers to the price index for a particular category of goods or services. **Relative price** is the price of a specific good or service in comparison with the prices of other goods and services. Good examples of relative prices include the prices for food and energy. The movements in a subindex or a relative price may be difficult to detect in the headline inflation rate. Because macroeconomic policy decision makers rely heavily on the headline inflation rate, they may not be aware of price movements at the subindex level. These prices movements, however, can be very useful for analyzing the prospects of an industry or a company. For example, if the producer price index for the machinery used by an industry rises quickly, the allowable capital depreciation permitted by the existing tax code may not generate sufficient tax benefits for the companies in that industry to meet future replacement expenses. The future profitability of the industry may decline for this reason. The decline in prices for flat-screen television sets provides an example of relative price movements. The price drop for these TVs may help to lower inflation pressure but can hurt manufacturers' profits.

EXAMPLE 6-9 Inflation

1. Which one of the following economic phenomena related to inflation cannot be determined by using observations of the inflation rate alone?
   A. Deflation
   B. Stagflation
   C. Hyperinflation
2. If a price index is calculated based on a fixed basket of goods, in an inflationary environment the inflation rate calculated based on this index over time will:
   A. overstate the actual cost of living.
   B. understate the actual cost of living.
   C. track the actual cost of living quite closely.

Solution to 1: B is correct. A high inflation rate alone does not indicate stagflation, which happens if high unemployment occurs together with high inflation.

Solution to 2: A is correct. The upward biases, such as the substitution bias and the quality bias, will overstate the actual cost of living.

4.2.4. Explaining Inflation

Economists describe two types of inflation: **cost-push**, in which rising costs, usually wages, compel businesses to raise prices generally; and **demand-pull**, in which increasing demand raise prices generally, which then are reflected in a business's costs as workers demand wage hikes to catch up with the rising cost of living. Whatever the sequence by which prices and costs rise in an economy, the fundamental cause is the same: excessive demands—for either raw materials, finished goods, or labor—that outstrip the economy's ability to respond. The initial signs appear in the areas with the greatest constraints: the labor market, the commodity market, or in some area of final output. Even before examining particular cost and price measures, practitioners, when considering inflation, look to indicators that might reveal when the economy faces such constraints.

4.2.4.1. Cost-Push Inflation In the area of cost-push or wage-push inflation, analysts can look for signs in commodity prices because commodities are an input to production. But because wages are the single biggest cost to businesses, they focus most particularly on the labor market. Because the object is to gauge demand relative to capacity, the unemployment rate is key, as well as measures of the number of workers available to meet the economy's expanding needs. Obviously, the higher the unemployment rate, the lower the likelihood that shortages will develop in labor markets, whereas the lower the unemployment rate, the greater the likelihood that shortages will drive up wages. Because the unemployment rate generally counts only people who are looking for work, some practitioners argue that it fails to account for the economy's full labor potential, and they state that a tight labor market will bring people out in search of work and ease any potential wage strains. To account for this issue and to modify the unemployment rate indicator, these practitioners also look at the participation rate

of people in the workforce, arguing that it gives a fuller and more accurate picture of potential than the unemployment rate.

Analysis in this area recognizes that not all labor is alike. Structural factors related to training deficiencies, cultural patterns in all or some of the population, inefficiencies in the labor market, and the like can mean that the economy will effectively face labor shortages long before the unemployment rate declines to very low figures. This effective unemployment rate, below which pressure emerges in labor markets, is frequently referred to as the **non-accelerating inflation rate of unemployment** (NAIRU) or, drawing on the work of the Nobel Prize winner Milton Friedman, the **natural rate of unemployment** (NARU). Of course, these rates vary from one economy to another and over time in a single economy. It is this rate rather than full employment that determines when an economy will experience bottlenecks in the labor market and wage-push inflationary pressures.

Take, for example, the technology sector. It has grown so rapidly in some economies that training in the workforce cannot keep up with demand. This sector can, as a consequence, face shortages of trained workers and attendant wage pressures even though the economy as a whole seems to have considerable slack in the overall labor market. Until training (supply) catches up with demand, that economy may carry a high NARU and NAIRU, yet experience wage and inflation pressure at rates of unemployment that in other places and circumstances might suggest ample slack in the labor market and much less wage-push pressure.

Of course, such assessments of wage-push inflation also find indicators in direct observations of the wage trends that, when they accelerate, might force businesses to raise prices (initiating the wage-price spiral mentioned earlier in this chapter). Statistical agencies provide a wide array of wage-cost indicators, such as hourly wage gauges, weekly earnings, and overall labor costs, including the outlays for benefits. Some of these indicators include the effects of special overtime pay or bonuses, whereas others do not. And although these measures give an idea of the cost to businesses and hence the kind of wage-push inflationary pressure, a complete picture emerges only when practitioners examine such trends alongside productivity measures.

Productivity, or output per hour, is an essential part of this inflation analysis because the output available from each worker determines the number of units over which businesses can spread the cost of worker compensation. The greater each worker's output is per hour, the lower the price businesses need to charge for each unit of output to cover hourly labor costs. And by extension, the faster output per hour grows, the faster labor compensation can expand without putting undue pressure on businesses' costs per unit of output. The equation for this **unit labor cost** (ULC) indicator, as it is called, is as follows:

$$\text{ULC} = W/O$$

where:

ULC = unit labor cost
W = total labor compensation per hour per worker
O = output per hour per worker

Many factors can affect labor productivity across time and between economies. The cyclical swings have already been described, as have the effects of technology and training. The pace of development also tends to increase worker productivity because the more sophisticated equipment, systems, and technologies workers have at their disposal, the higher their output per hour. Whatever causes the productivity growth, if it fails to keep up with

worker compensation, unit costs to a business rise and, as a business tries to protect its profit margins, prices generally come under increasing upward pressure. Generally this situation occurs because heavy demand for labor relative to available labor resources has pushed up compensation faster than productivity. Practitioners look for this relationship in the mix of indicators to identify cost- or wage-push inflationary pressure.

EXAMPLE 6-10 Unemployment Too High

1. Which of the following is *not* a problem with NARU and NAIRU?
   A. They work only in monetarist models.
   B. They may change over time given changes in technology and economic structure.
   C. They do not account for bottlenecks in segments of the labor market (e.g., college graduates).

Solution: A is correct. NARU and NAIRU are the unemployment rates at which the inflation rate will not rise because of a shortage of labor. This concept does not tie to a particular school of macroeconomic models.

4.2.4.2. Demand-Pull Inflation The search for indicators from the demand-pull side of the inflation question brings financial analysts back to the relationship between actual and potential real GDP and industrial capacity utilization. In a manner entirely analogous to the unemployment rate in the labor market, the higher the rate of capacity utilization or the closer actual GDP is to potential, the more likely an economy will suffer shortages, bottlenecks, a general inability to satisfy demand, and hence price increases—initially in commodities but ultimately more generally. And, of course, the more an economy operates below its potential or the lower the rate of capacity utilization, the less such supply pressure will exist and the greater likelihood of a slowdown in inflation, or outright deflation. In addition to these macroeconomic indicators, practitioners will also look for signs of inflationary pressure in commodity prices, in part because they are a cost to business, but more as a general sign of excess demand. For an individual economy, such observations could be misleading, however, because commodities trade in a global market and accordingly reflect global economic conditions more than those in an individual economy.

From an entirely different perspective, monetarists contend that inflation is fundamentally a monetary phenomenon. A surplus of money, they argue, will inflate the money price of everything in the economy. Stated in terms of straightforward supply and demand relationships, a surplus of money would bring down its value just as a surplus in any market would bring down the price of the product in excess. Because the price of money is stated in terms of the products it can buy, its declining value would have an expression in higher prices generally—that is, in inflation. This monetarist argument, as it is called, finds a simpler expression in the old saying that inflation results when too much money chases too few goods. Although it seems distant from other explanations of inflation, in practice it is not that

distinct. The excess of money creates the inflation by increasing liquidity, which ultimately causes a rapid rise in demand. In this sense, the monetarist argument is a special case under the more general heading of demand-pull concepts of inflation. The practical distinction between the monetarist and other approaches is in identifying the initial cause of the demand excess.

Financial analysts can track this effect by examining various money supply indicators, usually provided by the central bank. To detect an inflationary potential or the opposite, they note accelerations or decelerations in money growth from past trends. Obviously, accelerations, in the absence of a special explanation, signal the potential for inflationary pressure. In applying this approach, practitioners also compare money growth with the growth of the nominal economy, represented by nominal GDP. If money growth outpaces the growth of the nominal economy, there is an inflationary potential, especially if money growth has also accelerated from its trend. There is a disinflationary or deflationary potential if money growth trails the economy's rate of expansion, especially if it has also decelerated from its trend.

Velocity of Money (I)

Some practitioners view the likelihood of inflationary pressure from the vantage point of the ratio of nominal GDP to money supply, commonly called the "velocity of money." If this ratio remains stable around a constant or historical trend, they see reason to look for relative price stability. If velocity falls, it could suggest a surplus of money that might have inflationary potential, but much depends on why it has declined. If velocity has fallen because a cyclical correction has brought down the GDP numerator relative to the money denominator, then practitioners view prospects as more likely to lead to a cyclical upswing to reestablish the former relationship than inflationary pressure. If velocity has fallen, however, because of an increase in the money denominator, then inflationary pressure becomes more likely. If velocity rises, financial analysts might be concerned about a shortage of money in the economy and disinflation or deflation.

The 2008–2009 global recession and financial crisis offers an extreme example of these velocity ambiguities. As the global economy slipped into recession, which held back the GDP numerator in velocity measures, central banks, most notably the Federal Reserve in the United States, tried to help financial institutions cope by injecting huge amounts of money into their respective financial systems, raising the velocity denominator. Velocity measures plummeted accordingly. The expectation is that subsequent GDP growth as economies and financial markets heal will bring velocity back to a more normal level and trend. That said, the fear is that the monetary surge will, over the very long run, lead to inflation. For policy makers, this situation has created a very difficult policy choice. On one side, they need to sustain the supply of money to help their respective economies cope with the aftereffects of the financial crisis. On the other side, they need ultimately to withdraw any monetary excess to preclude potential inflationary pressures.

4.2.5. Inflation Expectations

Beyond demand-pull, monetary, and cost-push inflation considerations, practitioners also need to account for the effect of inflation expectations. Once inflation becomes embedded in an economy, businesses, workers, consumers, and economic actors of every kind begin to expect it and build those expectations into their actions. This reaction, in turn, creates an inflationary momentum of its own in a manner much like the wage-price spiral mentioned earlier in the chapter. Such expectations give inflation something of a self-sustaining character and cause it to persist in an economy even after its initial cause has disappeared. High inflation rates persisted in the 1970s and early 1980s in Europe and the United States on the basis of expectations even after these economies had sunk into recession. The resulting slow or negative economic growth combined with high unemployment and rising inflation was termed stagflation.

Measuring inflation expectations is not easy. Some practitioners gauge expectations by relying on past inflation trends and on the assumption that market participants largely extrapolate their past experiences. In some markets, surveys of inflation expectations are available, although these are often biased by the way the questions are asked. A third indicator becomes available when governments issue bonds that adjust in various ways to compensate holders for inflation, such as TIPS. By comparing the interest available on these bonds with other government bonds that do not offer such inflation-linked adjustments, practitioners can gauge the general level of inflation expectations among market participants and factor it into their own inflation forecasts and strategies.

For example, if today's yield on the 10-year nominal bond of a certain country is 3.5 percent and the yield on the 10-year inflation-protected bond of the same country is 1.5 percent, we infer that the market is pricing in a $3.5 - 1.5 = 2$ percent average annual inflation over the next 10 years. However, this calculation needs to be treated cautiously because the market for inflation-linked bonds is relatively small and thus yields can be influenced by other market factors. For example, in the past decade TIPS yields appear to have been artificially depressed by very strong demand from pension funds trying to match their liabilities at any cost.

EXAMPLE 6-11 Velocity of Money (II)

1. Cost-push inflation most likely occurs when:
   A. unemployment rates are low.
   B. unemployment rates are high.
   C. unemployment is either high or low.
2. Unit labor costs measure:
   A. hourly wage rates.
   B. total labor compensation per hour.
   C. a combination of hourly wages and output.
3. Demand-pull inflation:
   A. is a discredited concept.
   B. depends on the movements in commodity prices.
   C. reflects the state of economic activity relative to potential.

4. Monetarists believe inflation reflects:
   A. the growth of money.
   B. the level of interest rates.
   C. that there is no difference between monetarist positions and cost-push inflation.
5. The inflationary potential of a particular inflation rate depends on the economy's NAIRU or NARU, which in turn depends in part on:
   A. the intensity of past cyclical swings.
   B. the bargaining power of trade unions.
   C. the skill set of the workforce relative to the economy's industrial mix.
6. Which of the following is *not* a problem with NARU and NAIRU?
   A. They are not observable directly.
   B. They work only in monetarist models.
   C. They change over time given changes in technology and economic structure.

Solution to 1: A is correct. When unemployment is below NAIRU, there is a shortage of labor that pushes up labor cost.

Solution to 2: C is correct. Unit labor costs reflect the labor cost in each unit of output.

Solution to 3: C is correct. When the economy is operating above its potential capacity allowed by the resources available, inflation will start to rise.

Solution to 4: A is correct. Monetarists emphasize the role of money growth in determining the inflation rate, especially in the long run. As Milton Friedman famously put it: "Inflation is always and everywhere a monetary phenomenon."

Solution to 5: C is correct. If the skill set of a large part of the workforce cannot satisfy the hiring need from the employers, the NAIRU of such an economy can be quite high.

Solution to 6: B is correct. NAIRU and NARU reflect the potential of an economy and thus cannot be directly observed from the economic data. They also change over time depending on technological progress and social factors.

5. ECONOMIC INDICATORS

As used in business cycle contexts, an **economic indicator** is a variable that provides information on the state of the overall economy. Economic indicators are often classified according to whether they lag, lead, or coincide with changes in an economy's growth. **Leading economic indicators** have turning points that usually precede those of the overall economy. They are believed to have value for predicting the economy's future state, usually the near-term state. **Coincident economic indicators** have turning points that are usually close to those of the overall economy. They are believed to have value for identifying the economy's present state. **Lagging economic indicators** have turning points that take place later than those of the overall economy. They are believed to have value in identifying the economy's past condition.

To get as clear of a picture as possible, practitioners frequently consider several related indicators simultaneously. What follows is a review of these indicators and how practitioners use them.

5.1. Popular Economic Indicators

A very useful approach for practitioners is to take an aggregate perspective on leading, lagging, and coincident indicators. These aggregate measures typically are a composite of economic indicators known respectively to lead the cycle, run coincident with it, or lag it at cyclical turns. For obvious reasons, the leading indicators in particular help with anticipating cyclical turns up or down and allow strategists and others to position themselves and their companies in a secure and timely way to benefit from movements in the economic cycle.

The exact indicators combined into these composites vary from one economy to another. Even within an economy, they can have a remarkably diverse and eclectic character. In the United States, for instance, the composite leading indicator known as the **index of leading economic indicators** (LEI) has 10 component parts that run the gamut from orders for capital goods, to changes in the money supply, to swings in stock prices. Such composite indicators in other countries include equally eclectic combinations.

Similar statistics are available for numerous economies. The Conference Board, a U.S. industry research organization, computes leading, lagging, and coincident indicators for the United States and nine other countries plus the euro area (Eurozone). For about 30 countries and several aggregates, such as the EU and G-7, the Organization for Economic Cooperation and Development (OECD) calculates composite leading indicators (CLI) indexes, which gauge the state of the business cycle in the economy. One of the interesting features of CLI indexes is that they are consistent across countries, and therefore can be compared more easily to see how each region is faring. The Economic Cycle Research Institute (ECRI), a private company, also computes leading indicator indexes for about 20 countries on a weekly basis.

Although specifics for leading, coincident, and lagging indicators vary from one economy to another, they have much in common. In each case, they bring together various economic and financial measures that have displayed a consistently leading, coincident, or lagging relationship to that economy's general cycle. However, as reported by the Conference Board, the timing record of the various composite indexes for the United States has varied over the past 50 years. The coincident index closely matches the National Bureau of Economic Research (NBER) peak and trough dates, with eight of the past 13 turning points corresponding to the beginning or end of a recession. The leading indicator index displays more variability, leading cyclical contractions by eight to 20 months and expansions by one to 10 months.[10]

Exhibit 6-7 presents the 10 leading, four coincident, and seven lagging indicators tracked for the United States by the Conference Board. In addition to naming the indicators, it also offers a general description of why each measure fits in each of the three groups.

Let us consider a few examples that show the use of these statistics in identifying a business cycle phase. An increase in the reported ratio of consumer installment debt to income lags (occurs after) cyclical upturns; so the increase, by itself, would be evidence that an upturn has been underway. That could confirm the implication of positive changes in coincident indicators that an expansion is in place. As a leading economic indicator, a positive change in the S&P 500 index is supposed to lead (come before) an increase in aggregate economic activity. An increase in the S&P 500 would be positive for future economic growth, all else being equal. However, if the S&P 500 shows an increase but the aggregate index does not, we would likely not draw a positive conclusion. For a final example, if we observe that the LEI moved up a small amount on two consecutive observations, we might conclude that a modest economic expansion is expected.

[10]The Conference Board, *Business Cycle Indicators Handbook* (2001): 14, 15.

EXHIBIT 6-7 Leading, Coincident, and Lagging Indicators—United States

Indicator and Description	Reason
Leading	
1. Average weekly hours, manufacturing	Because businesses will cut overtime before laying off workers in a downturn and increase it before rehiring in a cyclical upturn, these measures move up and down before the general economy.
2. Average weekly initial claims for unemployment insurance	This measure offers a very sensitive test of initial layoffs and rehiring.
3. Manufacturers' new orders for consumer goods and materials	Because businesses cannot wait too long to meet demands for consumer goods or materials without ordering, these gauges tend to lead at upturns and downturns. Indirectly, they capture changes in business sentiment as well, which also often leads the cycle.
4. Vendor performance, slower deliveries diffusion index[a]	By measuring the speed at which businesses can complete and deliver an order, this gauge offers a clear signal of unfolding demands on businesses.
5. Manufacturers' new orders for nondefense capital goods	In addition to offering a first signal of movement, up or down, in an important economic sector, movement in this area also indirectly captures business expectations.
6. Building permits for new private housing units	Because most localities require permits before new building can begin, this gauge foretells new construction activity.
7. S&P 500 stock index	Because stock prices anticipate economic turning points, both up and down, their movements offer a useful early signal on economic cycles.
8. Money supply, real M2	Because money supply growth measures the tightness or looseness of monetary policy, increases in money beyond inflation indicate easy monetary conditions and a positive economic response, whereas declines in real M2 indicate monetary restraint and a negative economic response.
9. Interest rate spread between 10-year Treasury yields and overnight borrowing rates (federal funds rate)	Because long-term yields express market expectations about the direction of short-term interest rates, and rates ultimately follow the economic cycle up and down, a wider spread, by anticipating short-term rate increases, also anticipates an economic upswing. Conversely, a narrower spread, by anticipating short-term rate decreases, also anticipates an economic downturn.
10. Index of Consumer Expectations, University of Michigan	Because consumers make up about two-thirds of the U.S. economy and will spend more or less freely according to their expectations, this gauge offers early insight into future consumer spending and consequently directions in the whole economy.
Coincident	
1. Employees on nonfarm payrolls	Once recession or recovery is clear, businesses adjust their full-time payrolls.

(*Continued*)

EXHIBIT 6-7 *Continued*

Indicator and Description	Reason
2. Aggregate real personal income (less transfer payments)	By measuring the income flow from noncorporate profits and wages, this measure captures the current state of the economy.
3. Industrial Production Index	This index measures industrial output, thus capturing the behavior of the most volatile part of the economy. The service sector tends to be more stable.
4. Manufacturing and trade sales	In the same way as aggregate personal income and the industrial production index, this aggregate offers a measure of the current state of business activity.
Lagging	
1. Average duration of unemployment	Because businesses wait until downturns look genuine to lay off workers and wait until recoveries look secure to rehire, this measure is important because it lags the cycle on both the way down and the way up.
2. Inventory-to-sales ratio	Because inventories accumulate as sales initially decline and then, once a business adjusts its ordering, become depleted as sales pick up, this ratio tends to lag the cycle.
3. Change in unit labor costs	Because businesses are slow to fire workers, these costs tend to rise into the early stages of recession as the existing workforce is used less intensely. Late in the recovery when the labor market gets tight, upward pressure on wages can also raise such costs. In both cases, there is a clear lag at cyclical turns.
4. Average bank prime lending rate	Because this is a bank-administered rate, it tends to lag other rates that move either before cyclical turns or with them.
5. Commercial and industrial loans outstanding	Because these loans frequently support inventory building, they lag the cycle for much the same reason that the inventory-to-sales ratio does.
6. Ratio of consumer installment debt to income	Because consumers borrow heavily only when confident, this measure lags the cyclical upturn, but debt also overstays cyclical downturns because households have trouble adjusting to income losses, causing this ratio to lag in the downturn.
7. Change in consumer price index for services	Inflation generally adjusts to the cycle late, especially the more stable services area.

[a]A diffusion index usually measures the percentage of components in a series that are rising in the same period. It indicates how widespread a particular movement in the trend is among the individual components.

The component indicators for other countries, though different in specifics, are similar in most respects. The Eurozone, for instance, composes its leading index from eight components:

1. Economic sentiment index.
2. Residential building permits.

3. Capital goods orders.
4. The Euro Stoxx Equity Index.
5. M2 money supply.
6. An interest rate spread.
7. Eurozone Manufacturing Purchasing Managers Index.
8. Eurozone Service Sector Future Business Activity Expectations Index.

The parallels between many of these components and those used in the United States are clear, but Europe has a services component in its business activity measures that the United States lacks, and Europe forgoes many of the overtime and employment gauges that the United States includes.

Japan's leading index contains 10 components:

1. New orders for machinery and construction equipment.
2. Real operating profits.
3. Overtime worked.
4. Dwelling units started.
5. Six-month growth rate in labor productivity.
6. Business failures.
7. Business confidence (Tankan Survey).
8. Stock prices.
9. Real M2 money supply.
10. Interest rate spread.

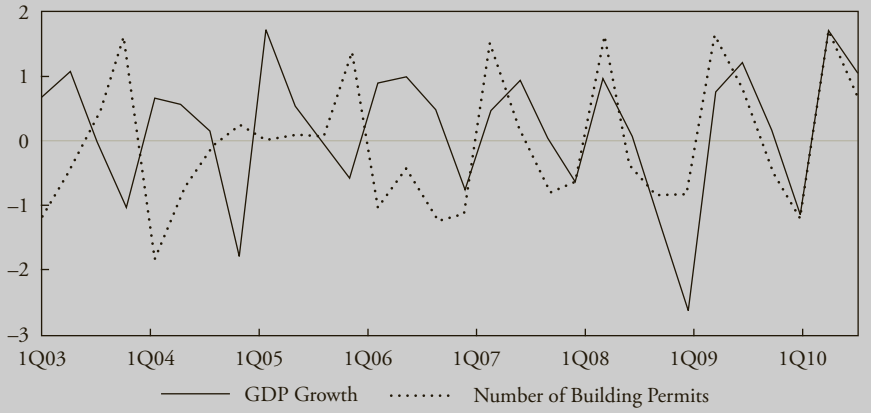
Again many are similar, but Japan includes labor market indicators more like the United States than Europe and adds a measure of business failures not included in the other two leading indexes.

Similarities and differences along these lines appear in indicators for the United Kingdom, Australia, South Africa, specific European economies, and other countries. The general tone is, however, similar to the detail provided here for the United States.

Building Permits as a Leading Economic Indicator

Exhibit 6-8 shows an example of a leading economic indicator in Germany, the granted building permits, along with its relationship to the growth of Germany's GDP. In Exhibit 6-8, the number of building permits usually peaks one quarter ahead of the GDP growth rate, except in 2008 and 2010. Before 2006, the growth rate of building permits usually bottomed out earlier than the GDP growth rate by four quarters. But after 2006, the troughs of the two series almost coincide. The uncertainty of the relationships between an indicator and business cycles is very common. Some indicators may be good predictors for economic expansions but poor predictors for recessions. This uncertainty is why economists and statisticians often combine different indicators and try to find common factors among them when building indicator indexes.

EXHIBIT 6-8 The Growth Rates of Germany GDP and Number of Building Permits



Note: The quarter-to-quarter growth rates are normalized by using the standard deviations of the two series, respectively.
Source: Federal Statistical Office of Germany.

Diffusion Index of Economic Indicators

In the United States, the Conference Board also compiles a monthly diffusion index of the leading, lagging, and coincident indicators. The **diffusion index** reflects the proportion of the index's components that are moving in a pattern consistent with the overall index. Analysts often rely on these diffusion indexes to provide a measure of the breadth of the change in a composite index.

For example, the Conference Board tracks the growth of each of the 10 constituents of its leading indicator measure, assigning a value of 1.0 to each indicator that rises by 0.05 percent or more during the monthly measurement period, a value of 0.5 for each component indicator that changes by less than 0.05 percent, and a value of 0 for each component indicator that falls by 0.05 percent or more. These assigned values, which of course differ in other indexes in other countries, are then summed and divided by 10 (the number of components). Then to make the overall measure resemble the more familiar indexes, the Board multiplies the result by 100.

A simple numerical example will help explain. Say, for ease of exposition, the indicator has only four component parts: stock prices, money growth, orders, and consumer confidence. In one month, stock prices rise 2.0 percent, money growth rises

1.0 percent, orders are flat, and consumer confidence falls by 0.6 percent. Using the Conference Board's assigned values, these would contribute respectively: 1.0 + 1.0 + 0.5 + 0 to create a numerator of 2.5. When divided by four (the number of components) and multiplied by 100, it generates an indicator of 62.5 for that month.

Assume that the following month stock prices fall 0.8 percent, money grows by 0.5 percent, orders pick up 0.5 percent, and consumer confidence grows 3.5 percent. Applying the appropriate values, the components would add to 0 + 1.0 + 1.0 + 1.0 = 3.0. Divided by the number of components and multiplied by 100, this yields an index value of 75. The 20.0 percent increase in the index value means more components of the composite index are rising. Given this result, an analyst can be more confident that the higher composite index value actually represents broader movements in the economy. In general, a diffusion index does not reflect outliers in any component (like a straight arithmetic mean would do) but instead tries to capture the overall change common to all components.

5.2. Other Variables Used as Economic Indicators

In addition to this array of measures, public agencies and trade associations provide aggregate cyclical measures. These may include surveys of industrialists, bankers, labor associations, and households on the state of their finances, level of activity, and their confidence in the future. In the United States, for instance, the Federal Reserve polls its 12 branches for a qualitative report on business activity and expectations in their respective regions. It summarizes those findings in what it calls the *Beige Book*, released every six weeks. Also in the United States, the Institute of Supply Management (ISM) polls its members to build indexes of manufacturing orders, output, employment, pricing, and comparable gauges for services. Over the past decade, so-called purchasing managers indexes along the lines of the ISM have been introduced in a wide range of countries, including Europe and China. Japan's industrial organization polls its members in a similar way and releases the findings in what is called the "Tankan Report." These diverse sources multiply within and across economies. Practitioners can use these sources to assess whether they confirm or contradict other more broad-based cyclical indicators, giving pause to, or greater confidence in, those earlier conclusions.

Using a statistical technique called principal components analysis, the Federal Reserve Bank of Chicago computes the Chicago Fed National Activity Index (CFNAI). The CFNAI is computed using 85 monthly macroeconomic series. These series cover industrial production, personal income, capital utilization, employment by sectors, housing starts, retail sales, and so on. Principal components analysis extracts the underlying trend that is common to most of these variables, thus distilling the essence of the U.S. business cycle. Similarly, the Bank of Italy in conjunction with the Centre for Economic Policy Research (CEPR) produces the EuroCOIN statistic, which is also based on principal components analysis. There are more than 100 macroeconomic series included in EuroCOIN. The EuroCOIN statistic also includes data derived from surveys, interest rates, and other financial variables. Both CFNAI and EuroCOIN are freely available online.

EXAMPLE 6-12 Economic Indicators

1. Leading, lagging, and coincident indicators are:
   A. the same worldwide.
   B. based on historical cyclical observations.
   C. based on Keynesian and/or monetarist theory.
2. A diffusion index:
   A. measures growth.
   B. reflects the consensus change in economic indicators.
   C. is roughly analogous to the indexes used to measure industrial production.
3. In the morning business news, a financial analyst learns that average hourly earnings increased last month. The most appropriate action for the analyst is to:
   A. call clients to inform them of a good trading opportunity today.
   B. examine other leading indicators to see any confirmation of a possible turning point for the economy.
   C. use the news in a research report as a confirmation for the belief that the economy has recovered from a recession.
4. Which one of the following is *not* thought to be a lagging indicator for the U.S. economy?
   A. Real M2
   B. Unit labor costs
   C. Commercial and industrial loans
5. The indicator indexes created by various organizations and research agencies:
   A. include only leading indicators to compute their value.
   B. are highly reliable signals on the phase of business cycles.
   C. evolve over time in terms of composition and computation formula.
6. Leading indicators are often very useful to investors because:
   A. they help investors to predict long-run returns on stocks.
   B. their turning points signal possible future change in the trend of economic growth.
   C. they may change into coincident indicators to help confirm the current state of the economy.

Solution to 1: B is correct. The recognition of economic indicators is based on empirical observations for an economy.

Solution to 2: B is correct. The diffusion indexes are constructed to reflect the common trends embedded in the movements of all the indicators included in such an index.

Solution to 3: B is correct. Financial analysts need to synthesize the information from various indicators in order to gather a reliable reading of the economic trends.

Solution to 4: A is correct. Real M2 is a leading indicator.

Solution to 5: C is correct. The indicator indexes are constantly updated for their composition and methodology based on the accumulation of empirical knowledge, and they can certainly include more than just leading indicators.

Solution to 6: B is correct. Leading indicators can help predict short-term trends but are not as useful in predicting long-run returns on equity. The changing nature of an indicator itself reduces its usefulness to investors.

6. SUMMARY

This chapter has summarized business cycle analysis. Among the points made are the following:

- Business cycles are a fundamental feature of market economies, but their amplitude and length vary considerably.
- Business cycles have four phases: trough, expansion, peak, and contraction.
- Keynesian theories focus on fluctuations of aggregate demand (AD). If AD shifts left, Keynesians advocate government intervention to restore full employment and avoid a deflationary spiral. Monetarists argue that the timing of government policies is uncertain and it is generally better to let the economy find its new equilibrium unassisted, but to ensure that the money supply is kept growing at an even pace.
- New classical and real business cycle (RBC) theories also consider fluctuations of aggregate supply (AS). If AS shifts left because of an input price increase or right because of a price decrease or technical progress, the economy will gradually converge to its new equilibrium. Government intervention is generally not necessary because it may exacerbate the fluctuation or delay the convergence to equilibrium. New Keynesians argue that frictions in the economy may prevent convergence, and government policies may be needed.
- The demand for factors of production may change in the short run as a result of changes in all components of GDP: consumption (e.g., households worry about the future, save more, and thus shift AD left); investment (e.g., companies expect customers to increase demand so they buy new equipment, thus shifting AD right; another example is that companies introduce new technologies, thus shifting long-term AS right); government (e.g., fiscal and monetary policies shift AD), and net exports (e.g., faster growth in other countries generates higher demand for the home country's products, thus shifting AD right, or higher prices of imported inputs shift AS left). Any shifts in AD and AS will affect the demand for the factors of production (capital and labor) that are used to produce the new level of GDP.
- Unemployment has different subcategories: frictional (people who are not working because they are between jobs); structural (people who are unemployed because they do not have the skills required by the openings or they reside far away from the jobs); discouraged workers, who are unemployed people who have given up looking for jobs because they do not believe they can find one (they are considered outside the labor force in unemployment statistics); and voluntarily unemployed people who do not wish to work, for example because they are in school, retired early, or are very rich (they are also considered outside the labor force in unemployment statistics).
- There are different types of inflation. Hyperinflation indicates a high (e.g., 100 percent annual) and increasing rate of inflation; deflation indicates a negative inflation rate (prices decrease); imported inflation is associated with increasing cost of inputs that come from

abroad; demand inflation is caused by constraints in production that prevent companies from making as many goods as the market demands (it is sometimes called wartime inflation because goods tend to be rationed in times of war).

- Economic indicators are statistics on macroeconomic variables that help in understanding which stage of the business cycle an economy is at. Of particular importance are the leading indicators, which suggest where the economy is likely to be in the near future. No economic indicator is perfect, and many of these statistics are subject to periodic revisions.
- Price levels are affected by real factors and monetary factors. Real factors include aggregate supply (an increase in supply leads to lower prices) and aggregate demand (an increase in demand leads to higher prices). Monetary factors include the supply of money (more money circulating, if the economy is in equilibrium, will lead to higher prices) and the velocity of money (higher velocity, if the economy is in equilibrium, will lead to higher prices).
- Inflation is measured by many indexes. Consumer price indexes reflect the prices of a basket of goods and services that is typically purchased by a normal household. Producer price indexes measure the cost of a basket of raw materials, intermediate inputs, and finished products. GDP deflators measure the price of the basket of goods and services produced within an economy in a given year. Core indexes exclude volatile items, such as agricultural products and energy, whose prices tend to vary more than other goods.

PRACTICE PROBLEMS[11]

1. Business cycle analysis *most* commonly describes economic activity that is conducted through:
   A. state enterprises.
   B. agricultural co-ops.
   C. private corporations.

2. The characteristic business cycle patterns of trough, expansion, peak, and contraction are:
   A. periodic.
   B. recurrent.
   C. of similar duration.

3. During the contraction phase of a business cycle, it is *most likely* that:
   A. inflation indicators are stable.
   B. aggregate economic activity is decreasing.
   C. investor preference for government securities declines.

4. An economic peak is *most* closely associated with:
   A. accelerating inflation.
   B. stable unemployment.
   C. declining capital spending.

[11]These practice problems were developed by Greg Gocek, CFA (Downers Grove, Illinois, USA).

5. Based on typical labor utilization patterns across the business cycle, productivity (output per hours worked) is *most likely* to be highest:
   A. at the peak of a boom.
   B. into a maturing expansion.
   C. at the bottom of a recession.

6. In a recession, companies are *most likely* to adjust their stock of physical capital by:
   A. selling it at fire sale prices.
   B. not maintaining equipment.
   C. quickly canceling construction activity.

7. The inventory-to-sales ratio is *most likely* to be rising:
   A. as a contraction unfolds.
   B. partially into a recovery.
   C. near the top of an economic cycle.

8. The Austrian economic school attributes the primary cause of the business cycle to:
   A. misguided government intervention.
   B. the creative destruction of technological progress.
   C. sticky price and wage expectations that exaggerate trends.

9. Monetarists favor a limited role for the government because they argue that:
   A. government policies operate with a lag.
   B. firms take time to adjust to systemic shocks to the economy.
   C. resource use is efficient with marginal revenue and cost equal.

10. The discouraged worker category is defined to include people who:
    A. are overqualified for their jobs.
    B. could look for a job but choose not to.
    C. currently look for work without finding it.

11. The unemployment rate is considered a lagging indicator because:
    A. new job types must be defined to count their workers.
    B. multiworker households change jobs at a slower pace.
    C. businesses are slow to hire and fire due to related costs.

12. The factor for which it is *most* difficult to estimate its effect on the unemployment rate is:
    A. technological progress.
    B. the use of temporary workers.
    C. the nature of underemployment.

13. The category of persons who would be *most likely* to be harmed by an increase in the rate of inflation is:
    A. homeowners with fixed 30-year mortgages.
    B. retirees relying on a fixed annuity payment.
    C. workers employed under contracts with escalator clauses.

14. The term that describes when inflation declines but nonetheless remains at a positive level is:
    A. deflation.
    B. stagflation.
    C. disinflation.

15. Deflation is *most likely* to be associated with:
    A. a shortage of government revenue.
    B. substantial macroeconomic contraction.
    C. explicit monetary policy to combat inflation.

16. The *least likely* consequence of a period of hyperinflation is the:
    A. reduced velocity of money.
    B. increased supply of money.
    C. possibility of social unrest.

The following information relates to Questions 17 and 18.

Consumption Baskets and Prices over Two Months

Date	November 2010		December 2010	
Goods	Quantity	Price	Quantity	Price
Sugar	70 kg	€0.90/kg	120 kg	€1.00/kg
Cotton	60 kg	€0.60/kg	50 kg	€0.80/kg

17. Assuming the base period for 2010 consumption is November and the initial price index is set at 100, then the inflation rate after calculating the December price index as a Laspeyres index is *closest* to:
    A. 19.2 percent.
    B. 36.4 percent.
    C. 61.6 percent.

18. For the December consumption basket, the value of the Paasche index is *closest* to:
    A. 116.
    B. 148.
    C. 160.

19. The characteristic of national consumer price indexes that is *most* typically shared across major economies worldwide is:
    A. the geographic areas covered in their surveys.
    B. the weights they place on covered goods and services.
    C. their use in the determination of macroeconomic policy.

20. Of the following statements regarding the producer price index (PPI), which is the *least likely*?
    A. The PPI can influence the future CPI.
    B. The PPI category weights can vary more widely than analogous CPI terms.
    C. The PPI is used more frequently than CPI as a benchmark for adjusting labor contract payments.

21. The inflation rate *most likely* relied on to determine public economic policy is:
    A. core inflation.
    B. headline inflation.
    C. an index of food and energy prices.

22. What is the *most* important effect of labor productivity in a cost-push inflation scenario?
    A. Rising productivity indicates a strong economy and a bias toward inflation.
    B. The productivity level determines the economy's status relative to its natural rate of unemployment.
    C. As productivity growth proportionately exceeds wage increases, product price increases are less likely.

23. Which of the following statements is the *best* description of the characteristics of economic indicators?
    A. Leading indicators are important because they track the entire economy.
    B. Lagging indicators in measuring past conditions do not require revisions.
    C. A combination of leading and coincident indicators can offer effective forecasts.

24. When the spread between 10-year U.S. Treasury yields and the federal funds rate narrows and at the same time the prime rate stays unchanged, this mix of indicators *most likely* forecasts future economic:
    A. growth.
    B. decline.
    C. stability.

25. If relative to prior values of their respective indicators, the inventory-to-sales ratio has risen, unit labor cost is stable, and real personal income has decreased, it is *most likely* that a peak in the business cycle:
    A. has occurred.
    B. is just about to occur.
    C. will occur sometime in the future.

CHAPTER 7

MONETARY AND FISCAL POLICY

Andrew Clare

Stephen Thomas

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Compare monetary policy and fiscal policy.
- Describe functions and definitions of money.
- Explain the money creation process.
- Describe theories of the demand for and supply of money.
- Describe the Fisher effect.
- Describe the roles and objectives of central banks.
- Contrast the costs of expected and unexpected inflation.
- Describe the implementation of monetary policy.
- Describe the qualities of effective central banks.
- Explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates.
- Contrast the use of inflation, interest rate, and exchange rate targeting by central banks.
- Determine whether a monetary policy is expansionary or contractionary.
- Describe the limitations of monetary policy.
- Describe the roles and objectives of fiscal policy.
- Describe the tools of fiscal policy, including their advantages and disadvantages.
- Describe the arguments for and against being concerned with the size of a fiscal deficit relative to gross domestic product (GDP).
- Explain the implementation of fiscal policy and the difficulties of implementation.
- Determine whether a fiscal policy is expansionary or contractionary.
- Explain the interaction of monetary policy and fiscal policy.

1. INTRODUCTION

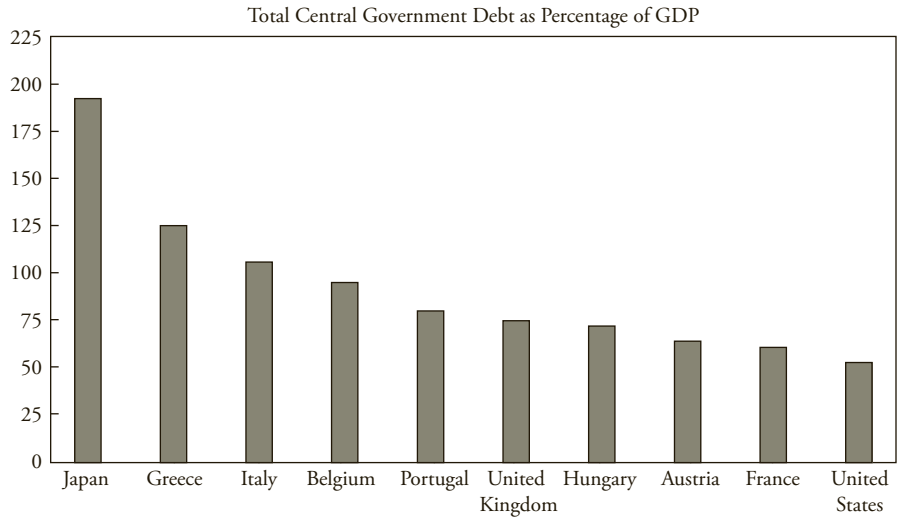
The economic decisions of households can have a significant impact on an economy. For example, decisions on the part of households to consume more and to save less can lead to an increase in employment, investment, and ultimately profits. Equally, the investment decisions made by corporations can have an important impact on the real economy and on corporate profits. But individual corporations can rarely affect large economies on their own; the decisions of a single household concerning consumption will have a negligible impact on the wider economy.

By contrast, the decisions made by governments can have an enormous impact on even the largest and most developed of economies for two main reasons. First, the public sectors of most developed economies normally employ a significant proportion of the population, and they are usually responsible for a significant proportion of spending in an economy. Second, governments are also the largest borrowers in world debt markets. Exhibit 7-1 gives some idea of the scale of government borrowing and spending.

Government policy is ultimately expressed through its borrowing and spending activities. In this chapter, we identify and discuss two types of government policy that can affect the macroeconomy and financial markets: monetary policy and fiscal policy.

Monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy.[1] By contrast, **fiscal policy** refers to the government's decisions about taxation and spending. Both monetary and fiscal policies are used to regulate economic activity over time. They can be used to accelerate growth when an economy starts to slow or to moderate growth and activity when an economy starts to overheat. In addition, fiscal policy can be used to redistribute income and wealth.

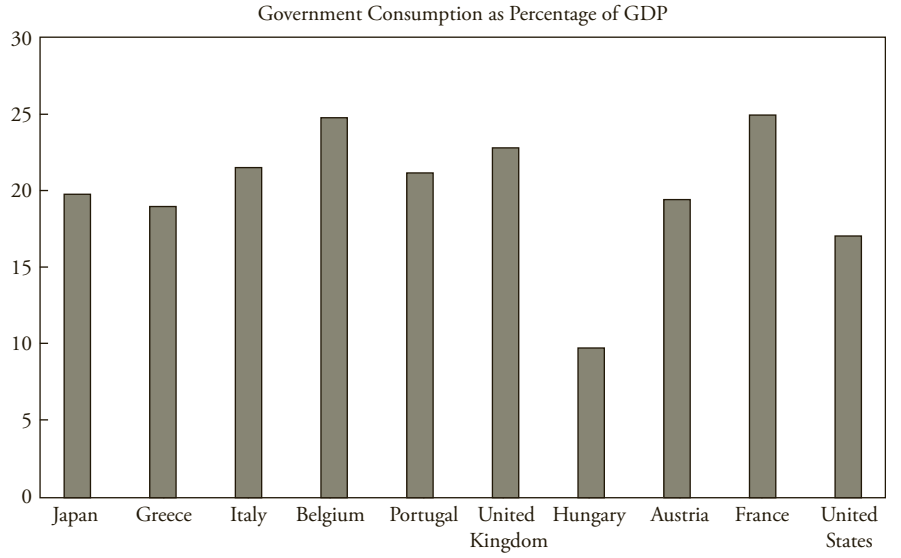
EXHIBIT 7-1a Central Government Debt to GDP, 2009



Source: Thomson Financial.

[1]Central banks can implement monetary policy almost completely independently of government interference and influence at one end of the scale, or simply as the agent of the government at the other end of the scale.

EXHIBIT 7-1b Public-Sector Spending to GDP, 2009



Source: Thomson Financial.

The overarching goal of both monetary and fiscal policy is normally the creation of an economic environment where growth is stable and positive and inflation is stable and low. Crucially, the aim is therefore to steer the underlying economy so that it does not experience economic booms that may be followed by extended periods of low or negative growth and high levels of unemployment. In such a stable economic environment, householders can feel secure in their consumption and saving decisions, while corporations can concentrate on their business and investment decisions, on making their regular coupon payments to their bondholders, and on making profits for their shareholders.

The challenges to achieving this overarching goal are many. Not only are economies frequently buffeted by shocks (such as oil price jumps), but some economists believe that natural cycles in the economy also exist. Moreover, there are plenty of examples from history where government policies—either monetary, fiscal, or both—have exacerbated an economic expansion that eventually led to damaging consequences for the real economy, for financial markets, and for investors.

The balance of the chapter is organized as follows. Section 2 provides an introduction to monetary policy and related topics. Section 3 presents fiscal policy. The interactions between monetary policy and fiscal policy are the subject of Section 4. A summary and practice problems conclude the chapter.

2. MONETARY POLICY

As stated earlier, monetary policy refers to government or central bank activities that are directed toward influencing the quantity of money and credit in an economy. Before we can begin to understand how monetary policy is implemented, we must examine the functions and role of money. We can then explore the special role that central banks play in today's economies.

EXAMPLE 7-1 Monetary and Fiscal Policy

1. Which of the following statements *best* describes monetary policy? Monetary policy:
   A. involves the setting of medium-term targets for broad money aggregates.
   B. involves the manipulation by a central bank of the government's budget deficit.
   C. seeks to influence the macroeconomy by influencing the quantity of money and credit in the economy.
2. Which of the following statements *best* describes fiscal policy? Fiscal policy:
   A. is used by governments to redistribute wealth and incomes.
   B. is the attempt by governments to balance their budgets from one year to the next.
   C. involves the use of government spending and taxation to influence economic activity.

Solution to 1: C is correct. Choice A is incorrect because, although the setting of targets for monetary aggregates is a possible *tool* of monetary policy, monetary policy itself is concerned with influencing the overall, or macro, economy.

Solution to 2: C is correct. Note that governments may wish to use fiscal policy to redistribute incomes and balance their budgets, but the overriding goal of fiscal policy is usually to influence a broader range of economic activity.

2.1. Money

To understand the nature, role, and development of money in modern economies, it is useful to think about a world without money—where to purchase any good or service an individual would have to pay with another good or service. An economy where such economic agents as households, corporations, and even governments pay for goods and services in this way is known as a **barter economy**. There are many drawbacks to such an economy. First, the exchange of goods for other goods (or services) would require both economic agents in the transaction to want what the other is offering. This means that there has to be a **double coincidence of wants**. It might also be impossible to undertake transactions where the goods are indivisible—that is, where one agent wishes to buy a certain amount of another's goods, but that agent only has one indivisible unit of another good that is worth more than the good that the agent is trying to buy. Another problem occurs if economic agents do not wish to exchange all of their goods for other goods and services. This may not be a problem, however, when the goods they have to sell can be stored safely so that they retain their value for the future. But if these goods are perishable, they will not be able to store value for their owner. Finally, in a barter economy, there are many measures of value: the price of oranges in terms of pears; of pears in terms of bread; of bread in terms of milk; or of milk in terms of oranges. A barter economy has no common measure of value that would make multiple transactions simple.

2.1.1. The Functions of Money

The most generic definition of **money** is that it is any generally accepted medium of exchange. A **medium of exchange** is any asset that can be used to purchase goods and services or to

repay debts. Money can thus eliminate the debilitating double coincidence of the "wants" problem that exists in a barter economy. When this medium of exchange exists, a farmer wishing to sell wheat for wine does not need to identify a wine producer in search of wheat. Instead, the farmer can sell wheat to those who want wheat in exchange for money. The farmer can then exchange this money for wine with a wine producer, who in turn can exchange that money for the goods or services that the wine producer wants.

However, for money to act as this liberating medium of exchange, it must possess certain qualities. It must:

- Be readily acceptable.
- Have a known value.
- Be easily divisible.
- Have a high value relative to its weight.
- Be difficult to counterfeit.

The first two qualities are closely related; the medium of exchange will be acceptable only if it has a known value. If the medium of exchange has the third quality, then it can be used to purchase items of relatively little value and of relatively large value with equal ease. Having a high value relative to its weight is a practical convenience, meaning that people can carry around sufficient wealth for their transaction needs. Finally, if the medium of exchange can be counterfeited easily, then it would soon cease to have a value and would not be readily acceptable as a means of effecting transactions; in other words, it would not satisfy the first two qualities.

Given the qualities that money needs to have, it is clear why precious metals (particularly gold and silver) often fulfilled the role of medium of exchange in early societies, and as recently as the early part of the twentieth century. Precious metals were acceptable as a medium of exchange because they had a known value, were easily divisible, had a high value relative to their weight, and could not be easily counterfeited.

Thus, precious metals were capable of acting as a medium of exchange. But they also fulfilled two other useful functions that are essential for the characteristics of money. In a barter economy, it is difficult to store wealth from one year to the next when one's produce is perishable, or indeed, if it requires large warehouses in which to store it. Because precious metals like gold had a high value relative to their bulk and were not perishable, they could act as a **store of wealth**. However, their ability to act as a store of wealth depended not only on the fact that they did not perish physically over time, but also on the belief that others would always value precious metals. The value from year to year of precious metals depended on people's continued demand for them in ornaments, jewelry, and so on. For example, people were willing to use gold as a store of wealth because they believed that it would remain highly valued. However, if gold became less valuable to people relative to other goods and services year after year, it would not be able to fulfill its role as a **store of value**, and as such might also lose its status as a medium of exchange.

Another important characteristic of money is that it can be used as a universal unit of account. As such, it can create a single unitary **measure of value** for all goods and services. In an economy where gold and silver are the accepted medium of exchange, all prices, debts, and wealth can be recorded in terms of their gold or silver coin exchange value. Money, in its role as a unit of account, drastically reduces the number of prices in an economy compared to barter, which requires that prices be established for a good in terms of all other goods for which it might be exchanged.

In summary, money fulfills three important functions:

1. It acts as a medium of exchange.
2. It provides individuals with a way of storing wealth.
3. It provides society with a convenient measure of value and unit of account.

2.1.2. Paper Money and the Money Creation Process

Although precious metals like gold and silver fulfilled the required functions of money relatively well for many years, and although carrying gold coins around was easier than carrying around one's physical produce, it was not necessarily a safe way to conduct business.

A crucial development in the history of money was the **promissory note**. The process began when individuals began leaving their excess gold with goldsmiths, who would look after it for them. In turn the goldsmiths would give the depositors a receipt, stating how much gold they had deposited. Eventually these receipts were traded directly for goods and services, rather than there being a physical transfer of gold from the goods buyer to the goods seller. Of course, both the buyer and seller had to trust the goldsmith because the goldsmith had all the gold and the goldsmith's customers had only pieces of paper. These depository receipts represented a promise to pay a certain amount of gold on demand. This paper money therefore became a proxy for the precious metals on which they were based; that is, they were directly related to a physical commodity. Many of these early goldsmiths evolved into banks, taking in excess wealth and in turn issuing promissory notes that could be used in commerce.

In taking in other people's gold and issuing depository receipts and later promissory notes, it became clear to the goldsmiths and early banks that not all the gold that they held in their vaults would be withdrawn at any one time. Individuals were willing to buy and sell goods and services with the promissory notes, but the majority of the gold that backed the notes just sat in the vaults—although its ownership would change with the flow of commerce over time. A certain proportion of the gold that was not being withdrawn and used directly for commerce could therefore be lent to others at a rate of interest. By doing this, the early banks "created" money.

The process of **money creation** is a crucial concept for understanding the role that money plays in an economy. Its potency depends on the amount of money that banks keep in reserve to meet the withdrawals of its customers. This practice of lending customers' money to others on the assumption that not all customers will want all of their money back at any one time is known as **fractional reserve banking**.

We can illustrate how it works through a simple example. Suppose that the bankers in an economy come to the view that they need to retain only 10 percent of any money deposited with them. This is known as the **reserve requirement**.[2] Now consider what happens when a customer deposits €100 in the First Bank of Nations. This deposit changes the balance sheet of the First Bank of Nations, as shown in Exhibit 7-2, and it represents a liability to the bank because it is effectively lent to the bank by the customer. By lending 90 percent of this deposit to another customer the bank has two types of assets: (1) the bank's reserves of €10 and (2) the loan equivalent to €90. Notice that the balance sheet still balances; €100 worth of assets and €100 worth of liabilities are on the balance sheet.

Now suppose that the recipient of the loan of €90 uses this money to purchase some goods of this value and the seller of the goods deposits this €90 in another bank, the Second Bank of Nations. The Second Bank of Nations goes through the same process; it retains €9 in reserve and loans 90 percent of the deposit (€81) to another customer. This customer in turn spends €81 on some goods or services. The recipient of this money deposits it at the Third Bank of Nations, and so on. This example shows how money is created when a bank makes a loan.

[2]This is an example of a *voluntary* reserve requirement because it is self-imposed.

EXHIBIT 7-2 Money Creation via Fractional Reserve Banking

First Bank of Nations

Assets		Liabilities	
Reserves	€10	Deposits	€100
Loans	€90		

Second Bank of Nations

Assets		Liabilities	
Reserves	€9	Deposits	€90
Loans	€81		

Third Bank of Nations

Assets		Liabilities	
Reserves	€8.1	Deposits	€81
Loans	€72.9		

This process continues until there is no more money left to be deposited and lent out. The total amount of money "created" from this one deposit of €100 can be calculated as:

$$\text{New deposit/Reserve requirement} = €100/0.10 = €1{,}000 \qquad (7\text{-}1)$$

It is the sum of all the deposits now in the banking system. You should also note that the original deposit of €100, via the practice of reserve banking, was the catalyst for €1,000 worth of economic transactions. That is not to say that economic growth would be zero without this process, but instead that it can be an important component in economic activity.

The amount of money that the banking system creates through the practice of fractional reserve banking is a function of 1 divided by the reserve requirement, a quantity known as the **money multiplier**.[3] In the case just examined, the money multiplier is $1/0.10 = 10$. Equation 7-1 implies that the smaller the reserve requirement, the greater the money multiplier effect.

In our simplistic example, we assumed that the banks themselves set their own reserve requirements. However, in some economies, the central bank sets the reserve requirement, which is a potential means of affecting money growth. In any case, a prudent bank would be wise to have sufficient reserves such that the withdrawal demands of its depositors can be met in stressful economic and credit market conditions.

Later, when we discuss central banks and central bank policy, we will see how central banks can use the mechanism just described to affect the money supply. Specifically, the central bank could, by purchasing €100 in government securities credited to the bank account of the seller, seek to initiate an increase in the money supply. The central bank may also lend reserves directly to banks, creating excess reserves (relative to any imposed or self-imposed reserve requirement) that can support new loans and money expansion.

[3]This quantity, known as the simple money multiplier, represents a maximum expansion. To the extent that banks hold excess reserves or that money lent out is not redeposited, the money expansion would be less. More complex multipliers incorporating such factors are developed in more advanced texts.

EXAMPLE 7-2 Money and Money Creation

1. To fulfill its role as a medium of exchange, money should:
   A. be a conservative investment.
   B. have a low value relative to its weight.
   C. be easily divisible and a good store of value.
2. If the reserve requirement for banks in an economy is 5 percent, how much money could be created with the deposit of an additional £100 into a deposit account?
   A. £500
   B. £1,900
   C. £2,000
3. Which of the following functions does money normally fulfill for a society?
   A. Money acts as a medium of exchange only.
   B. It provides economic agents with a means of storing wealth only.
   C. It provides society with a unit of account, acts as a medium of exchange, and acts as a store of wealth.

Solution to 1: C is correct. Money needs to have a known value and be easily divisible. It should also be readily acceptable, difficult to counterfeit, and have a high value relative to its weight.

Solution to 2: C is correct. To calculate the increase in money from an additional deposit in the banking system, use the following expression: new deposit/reserve requirement.

Solution to 3: C is correct. Money needs to be able to fulfill the functions of acting as a unit of account, a medium of exchange, and a means of storing wealth.

2.1.3. Definitions of Money

The process of money creation raises a fundamental issue: What is money? In an economy with money but without promissory notes and fractional reserve banking, money is relatively easy to define: Money is the total amount of gold and silver coins in circulation, or their equivalent. The money creation process just described, however, indicates that a broader definition of money might encompass all the notes and coins in circulation *plus* all bank deposits.

More generally, we might define money as any medium that can be used to purchase goods and services. Notes and coins can be used to fulfill this purpose, and yet such currency is not the only means of purchasing goods and services. Personal checks can be written based on a bank checking account, while debit cards can be used for the same purpose. But what about time deposits or savings accounts? Nowadays transfers can be made relatively easily from a savings account to a current account; therefore, these savings accounts might also be considered as part of the stock of money. Credit cards are also used to pay for goods and services; however, there is an important difference between credit card payments and those made by checks and debit cards. Unlike a check or debit card payment, a credit card payment involves a deferred payment. Basically, the greater the complexity of any financial system, the harder it is to define money.

The monetary authorities in most modern economies produce a range of measures of money (see Exhibit 7-3). But generally speaking, the money stock consists of notes and coins in circulation, plus the deposits in banks and other financial institutions that can be readily

EXHIBIT 7-3 Definitions of Money

Money Measures in the United States

The U.S. Federal Reserve produces two measures of money. The first is M1, which comprises notes and coins in circulation, travelers' checks of nonbank issuers, demand deposits at commercial banks, and other deposits on which checks can be written. M2 is the broadest measure of money currently produced by the Federal Reserve and includes M1, plus savings and money market deposits, time deposit accounts of less than $100,000, and other balances in retail money market and mutual funds.

Money Measures in the Eurozone

The European Central Bank (ECB) produces three measures of euro area money supply. The narrowest is M1. M1 comprises notes and coins in circulation, plus all overnight deposits. M2 is a broader definition of euro area money that includes M1, plus deposits redeemable with notice up to three months and deposits with maturity up to two years. Finally, the euro area's broadest definition of money is M3, which includes M2, plus repurchase agreements, money market fund units, and debt securities with up to two years maturity.

Money Measures in Japan

The Bank of Japan calculates three measures of money. M1 is the narrowest measure and consists of cash currency in circulation. M2 incorporates M1 but also includes certificates of deposit (CDs). The broadest measure, M3, incorporates M2, plus deposits held at post offices, as well as other savings and deposits with financial institutions. There is also a "broad measure of liquidity" that encompasses M3 as well as a range of other liquid assets, such as government bonds and commercial paper.

Money Measures in the United Kingdom

The United Kingdom produces a set of four measures of the money stock. M0 is the narrowest measure and comprises notes and coins held outside the bank of England, plus bankers' deposits at the Bank of England. M2 includes M0, plus (effectively) all retail bank deposits. M4 includes M2, plus wholesale bank and building society deposits and also certificates of deposit. Finally, the Bank of England produces another measure called M3H, which is a measure created to be comparable with money definitions in the European Union. M3H includes M4, plus U.K. residents' and corporations' foreign currency deposits in banks and building societies.

used to make purchases of goods and services in the economy. In this regard, economists often speak of the rate of growth of **narrow money** and **broad money**. By narrow money, they generally mean the notes and coins in circulation in an economy, plus other very highly liquid deposits. Broad money encompasses narrow money but also includes the entire range of liquid assets that can be used to make purchases.

Because financial systems, practices, and institutions vary from economy to economy, so do definitions of money; thus, it is difficult to make international comparisons. Still, most central banks produce both a narrow and a broad measure of money, plus some intermediate ones, too. Exhibit 7-3 shows the money definitions in four economies.

2.1.4. The Quantity Theory of Money

The previous section of this chapter shows that there are many definitions of money. In this section, we explore the important relationship between money and the price level. This relationship is best expressed in the **quantity theory of money**, which asserts that total

spending (in money terms) is proportional to the quantity of money. The theory can be explained in terms of Equation 7-2, known as the **quantity equation of exchange**:

$$M \times V = P \times Y \tag{7-2}$$

where M is the quantity of money, V is the velocity of circulation of money (the average number of times in a given period that a unit of currency changes hands), P is the average price level, and Y is real output. The expression is really just an accounting identity. Effectively, it says that over a given period, the amount of money used to purchase all goods and services in an economy, $M \times V$, is equal to monetary value of this output, $P \times Y$. If the velocity of money is approximately constant—which is an assumption of quantity theory—then spending $P \times Y$ is approximately proportional to M. The quantity equation can also be used to explain a consequence of **money neutrality**. If money neutrality holds, then an increase in the money supply, M, will not affect Y, real output, or the speed with which money changes hands, V, because if real output is unaffected, there would be no need for money to change hands more rapidly.[4] However, it will cause the aggregate price level, P, to rise.

The simple quantity theory gave rise to the equally simple idea that the price level, or at least the rate of inflation, can be controlled by manipulating the rate of growth of the money supply. Economists who believe this are referred to as **monetarists**. They argue that there is a causal relationship running from money growth to inflation. In the past, some governments have tried to apply this logic in their efforts to control inflation, most notably and unsuccessfully the United Kingdom's government in 1979 (see upcoming box, "Mrs. Thatcher's Monetary Experiment"). However, it is possible that causality runs the other way—that is, from real activity to the money supply. This means that the quantity of money in circulation is determined by the level of economic activity, rather than vice versa.

2.1.5. The Demand for Money

The amount of wealth that the citizens of an economy choose to hold in the form of money—as opposed to bonds or equities—is known as the demand for money. There are three basic motives for holding money:

1. Transactions-related
2. Precautionary
3. Speculative

Money balances that are held to finance transactions are referred to as **transactions money balances**. The size of the transactions balances will tend to increase with the average value of transactions in an economy. Generally speaking, as gross domestic product (GDP) grows over time, transactions balances will also tend to grow; however, the ratio of transactions balances to GDP remains fairly stable over time.

As the name suggests, **precautionary money balances** are held to provide a buffer against unforeseen events that might require money. These balances will tend to be larger for individuals or organizations that enter into a high level of transactions over time. In other words, a

[4]Note that the full version of the quantity theory of money uses the symbol T rather than Y to indicate transactions, because money is used not just for buying goods and services but also for financial transactions. We will return to this point in the discussion of quantitative easing.

precautionary buffer of $100 for a company that regularly enters into transactions worth millions of dollars might be considered rather small. When we extend this logic to the overall economy, we can see that these precautionary balances will also tend to rise with the volume and value of transactions in the economy, and therefore, GDP as well.

Finally, the **speculative demand for money** (sometimes called the **portfolio demand for money**) relates to the demand to hold speculative money balances based on the potential opportunities or risks that are inherent in other financial instruments (e.g., bonds). **Speculative money balances** consist of monies held in anticipation that other assets will decline in value. But in choosing to hold speculative money balances rather than bonds, investors give up the return that could be earned from the bonds or other financial assets. Therefore, the speculative demand for money will tend to fall as the returns available on other financial assets rise. However, it will tend to rise as the perceived risk in other financial instruments rises. In equilibrium, individuals will tend to increase their holdings of money relative to riskier assets until the marginal benefit of having a lower-risk portfolio of wealth is equal to the marginal cost of giving up a unit of expected return on these riskier assets. In aggregate, then, speculative balances will tend to be inversely related to the expected return on other financial assets and directly related to the perceived risk of other financial assets.

EXAMPLE 7-3 Money

1. The transactions demand for money refers to the demand to hold money:
   A. as a buffer against unforeseen events.
   B. to use in the purchase of goods and services.
   C. based on the opportunity or risks available on other financial instruments.
2. The speculative demand for money will tend to:
   A. fall as the perceived risk on other assets rises.
   B. rise as the expected returns on other assets fall.
   C. be inversely related to the transactions demand for money.
3. To define the difference between narrow and broad money, broad money:
   A. is limited to those liquid assets most commonly used to make purchases.
   B. can be used to purchase a wider range of goods and services than narrow money.
   C. encompasses narrow money and refers to the stock of the entire range of liquid assets that can be used to make purchases.

Solution to 1: B is correct. The transactions demand for money refers to the amount of money that economic agents wish to hold to pay for goods and services.

Solution to 2: B is correct. If the expected return on other assets falls, then the opportunity cost of holding money also falls and can, in turn, lead to an increase in the speculative demand for money.

Solution to 3: C is correct. This is the definition of broad money. Broad money encompasses narrow money.

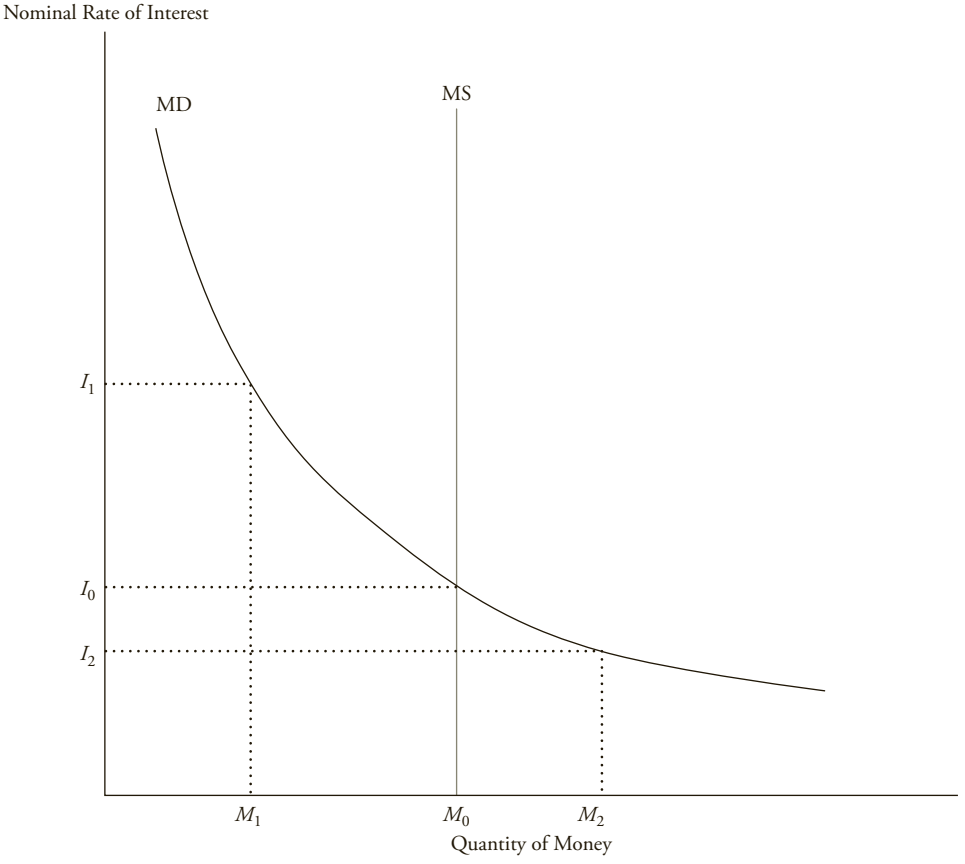
2.1.6. The Supply of and Demand for Money

We have now discussed definitions of money, its relationship with the aggregate price level, and the demand for it. We now discuss the interaction between the supply of and demand for money.

As with most other markets, the supply of money and the demand to hold it will interact to produce an equilibrium price for money. In this market, the price of money is the nominal interest rate that could be earned by lending it to others. Exhibit 7-4 shows the supply and demand curves for money. The vertical scale represents the rate of interest; the horizontal scale plots the quantity of nominal money in the economy. The supply curve (MS) is vertical because we assume that there is a fixed nominal amount of money circulating at any one time. The demand curve (MD) is downward sloping because as interest rates rise, the speculative demand for money falls. The supply of and demand for money are both satisfied at an equilibrium interest rate of I_0. I_0 is the rate of interest at which no excess money balances exist.

To see why I_0 is the equilibrium rate of interest where there are no excess money balances, consider the following. If the interest rate on bonds were I_1 instead of I_0, there would be excess supply of money ($M_0 - M_1$). Economic agents would seek to buy bonds with their

EXHIBIT 7-4 The Supply of and Demand for Money



excess money balances, which would force the price of bonds up and the interest rate back down to I_0. Similarly, if bonds offered a rate of interest I_2, there would be an excess demand for money ($M_2 - M_0$). Corporations and individuals would seek to sell bonds so that individuals could increase their money holdings, but in doing so, the price of bonds would fall and the interest rate offered on them would rise until it reached I_0. Interest rates effectively adjust to bring the market into equilibrium (clear the market). In this simple example, we have also assumed that the supply of money and bonds is fixed as economic agents readjust their holdings. In practice, this may not be true, but the dynamics of the adjustment process described here essentially still hold.

Exhibit 7-4 also reemphasizes the relationship between the supply of money and the aggregate price level, which we first encountered when discussing the quantity theory of money. Suppose that the central bank increases the supply of money from M_0 to M_2, so that the vertical supply curve shifts to the right. Because the increase in the supply of money makes it more plentiful and hence less valuable, its price (the interest rate) falls as the price level rises.

This all sounds very simple, but in practice the effects of an increase in the money supply are more complex. The initial increase in the money supply will create excess supply of cash. People and companies could get rid of the excess by lending the money to others by buying bonds, as implied earlier, but they might also deposit it in a bank or simply use it to buy goods and services. But an economy's capacity to produce goods and services depends on the availability of real things: notably, natural resources, capital, and labor—that is, factors of production supplied either directly or indirectly by households. Increasing the money supply does not change the availability of these real things. Thus, some economists believe that the long-run impact of an exogenous increase in the supply of money is an increase in the aggregate price level.

This phenomenon—whereby an increase in the money supply is thought in the long run simply to lead to an increase in the price level while leaving real variables like output and employment unaffected—is known as **money neutrality**. To see why in the long run money should have a neutral effect on real things, consider the following simple example.

Suppose the government declared today that 1 kilogram would henceforth be referred to as 2 kilograms and that 1.5 kilograms would be referred to as 3 kilograms. In other words, suppose the government halved the value of a kilogram. Would anything real have changed? A 1 kilogram bag of sugar would not have changed physically, although it would be relabeled as a 2 kilogram bag of sugar. However, there might be some short-run effects; confused people might buy too little sugar, and some people might go on crash diets! But ultimately people would adjust; in the long run, the change wouldn't matter. There is a clear parallel here with the theory of money neutrality. Doubling the prices of everything—halving the value of a currency—does not change anything real. This is because, like kilograms, money is a unit of account. However, halving the value of a currency could affect real things in the short run.

There are two points worth making with regard to money neutrality. First, although the simple kilogram analogy does suggest that money should not affect real things in the long run, as the British economist Keynes said: "In the long run we are all dead." In practice, it is very difficult for economists to be sure that money neutrality holds in the long run. And second, we must assume that monetary authorities do believe that the money supply can affect real things in the short run. If they did not, then there would be almost no point to monetary policy.

2.1.7. The Fisher Effect

The **Fisher effect** is directly related to the concept of money neutrality. Named after the economist Irving Fisher, the Fisher effect states that the real rate of interest in an economy is

stable over time so that changes in nominal interest rates are the result of changes in expected inflation. Thus, the nominal interest rate (R_{nom}) in an economy is the sum of the required real rate of interest (R_{real}) and the expected rate of inflation (π^e) over any given time horizon:

$$R_{nom} = R_{real} + \pi^e \tag{7-3}$$

According to money neutrality, over the long term the money supply and the growth rate in money should not affect R_{real} but will affect inflation and inflation expectations.

The Fisher effect also demonstrates that embedded in every nominal interest rate is an expectation of future inflation. Suppose that 12-month U.S. government Treasury bills offered a yield equal to 4 percent over the year. Suppose also that T-bill investors wished to earn a real rate of interest of 2 percent and expected inflation to be 2 percent over the next year. In this case, the return of 4 percent would be sufficient to deliver the investors' desired real return of 2 percent (so long as inflation did not exceed 2 percent). Now suppose that investors changed their view about future inflation and instead expected it to equal 3 percent over the next 12 months. To compensate them for the higher expected inflation, the T-bill rate would have to rise to 5 percent, thereby preserving the required 2 percent real return.

There is one caveat to this example. Investors can never be sure about future values of such economic variables as inflation and real growth. To compensate them for this uncertainty, they require a **risk premium**. The greater the uncertainty, the greater the required risk premium. So all nominal interest rates are actually comprised of three components:

1. A required real return.
2. A component compensating investors for expected inflation.
3. A risk premium to compensate them for uncertainty.

EXAMPLE 7-4 Interest Rates and the Supply of Money

1. According to the quantity equation of exchange, an increase in the money supply can lead to:
   A. an increase in the aggregate price level regardless of changes in the velocity of circulation of money.
   B. an increase in the aggregate price level as long as the velocity of circulation of money rises sufficiently to offset the increase in the money supply.
   C. an increase in the aggregate price level as long as the velocity of circulation of money does not fall sufficiently to offset the increase in the money supply and real output is unchanged.
2. The nominal interest rate comprises a real rate of interest:
   A. plus a risk premium only.
   B. plus a premium for expected inflation only.
   C. plus compensation for both expected inflation and risk.
3. An expansion in the money supply would *most likely*:
   A. lead to a decline in nominal interest rates.
   B. lead to an increase in nominal interest rates.
   C. reduce the equilibrium amount of money that economic agents would wish to hold.

Solution to 1: C is correct. If the velocity of circulation of money does not change with an increase in the money supply and real output is fixed, then the aggregate price level should increase. If the velocity of circulation of money falls sufficiently, or if real output rises sufficiently, then the increase in money may have no impact on prices.

Solution to 2: C is correct. Investors demand a real rate of interest plus compensation for expected inflation and a risk premium to compensate them for uncertainty.

Solution to 3: A is correct. Increasing the supply of money, all other things being equal, will reduce its price—that is, the interest rate on money balances.

Mrs. Thatcher's Monetary Experiment

The Background

Over the 1970s, the United Kingdom had one of the worst inflation records of any developed economy. Retail price inflation averaged 12.6 percent over that decade and peaked at 26.9 percent in August 1975. Over this period, Margaret Thatcher and her advisers became convinced that inflation could not be controlled by the income and price policies used in the United Kingdom in the past. Instead, they believed that inflation could be tamed by controlling the rate of growth of the money supply. Prime Minister Thatcher's first administration took power in May 1979 with the intention of pursuing a monetarist agenda—that is, a macroeconomic policy that would be underpinned by targets for money supply growth.

The Medium-Term Financial Strategy

Targets for monetary growth were set for a definition of the money supply known as Sterling M3 (£M3), which was to be kept in the range of 7 to 11 percent for the period 1980–1981 and then gradually reduced to within 4 to 8 percent by 1983–1984. This set of targets was known as the Medium-Term Financial Strategy (MTFS). The idea was simple: control the rate of growth of the money supply, and the rate of growth of prices (i.e., inflation) would remain under control, too. The instrument of control was the Bank of England's policy interest rate that would be set to achieve the desired rate of growth of the money supply. This was a macroeconomic policy built, however imperfectly, on an interpretation of the quantity theory of money.

The theory was simple, but the practice proved to be less so. Over the first two and a half years of the MTFS, £M3 overshot its target by 100 percent. The inability of the monetary authorities to control the rate of growth of the broad money supply was largely caused by Thatcher's abolition of exchange controls in 1979. Without these controls, there was a significant increase in foreign exchange business that came into the British banking system, which changed the velocity of money and therefore meant that the relationship between broad money and nominal incomes had changed fundamentally.*

Despite the inability to control the money supply, in 1983 the Thatcher administration reasserted its confidence in the policy and published a further set of monetary targets for several years ahead. However, the persistent failure to meet these targets, too, eventually led to the abandonment of any type of monetary targeting by the summer of 1985.

The experience of the U.K. monetary authorities over this period emphasizes how unstable the relationship between money and the policy interest rate could be, along with the relationship between money and aggregate demand—particularly in an economy experiencing rapid financial innovation, as the U.K. economy was following the abolition of exchange controls and the introduction of greater competition within the banking industry.

Today the Bank of England is responsible for the operation and implementation of monetary policy in the United Kingdom. The trends in money supply are watched very carefully, but they are not the subject of targets per se.

*See Goodhart (1989) for a discussion.

2.2. Roles of Central Banks

Central banks play a number of key roles in modern economies. Generally, a **central bank** is the monopoly supplier of the currency, the banker to the government and the bankers' bank, the lender of last resort, the regulator and supervisor of the payments system, the conductor of monetary policy, and the supervisor of the banking system. Let us examine these roles in turn.

In its earliest form, money could be exchanged for a prespecified precious commodity, usually gold, and promissory notes were issued by many private banks. Today, however, state-owned institutions—usually central banks—are designated in law as being the monopoly suppliers of a currency. Initially, these monopolists supplied money that could be converted into a prespecified amount of gold; they adhered to a **gold standard**. For example, up until 1931, bank notes issued by Britain's central bank, the Bank of England, could be redeemed at the bank for a prespecified amount of gold. But Britain, like most other major economies, abandoned this convertibility principle in the first half of the twentieth century. Money in all major economies today is not convertible by law into anything else, but it is, in law, **legal tender**. This means that it must be accepted when offered in exchange for goods and services. Money that is not convertible into any other commodity is known as **fiat money**. Fiat money derives its value via government decree and because people accept it for payment of goods and services and for debt repayment.

As long as fiat money is acceptable to everyone as a medium of exchange and it holds its value over time, then it will also be able to serve as a unit of account. However, once an economy has moved to a system of fiat money, the role of the supplier of that money becomes even more crucial because the authorities could, for example, expand the supply of this money indefinitely should they wish to do so. Central banks therefore play a crucial role in modern economies as the suppliers and guardians of the value of their fiat currencies and as institutions charged with the role of maintaining confidence in their currencies. As the monopoly suppliers of an economy's currency, central banks are at the center of economic life. As such, they assume other roles in addition to being the suppliers and guardians of the value of their currencies.

Most central banks act as the banker to the government and to other banks. They also act as a **lender of last resort** to banks. Because the central bank effectively has the capacity to print money, it is in the position to be able to supply the funds to banks that are facing a damaging shortage. The economic agents know that the central bank stands ready to provide the liquidity required by any of the banks under its jurisdiction and they trust government bank deposit insurance to help to prevent bank runs in the first place. However, the recent financial crisis has shown that this knowledge is not always sufficient to deter a bank run.

The Northern Rock Bank Run

In the latter part of the summer of 2007, the fall in U.S. house prices along with the related implosion of the U.S. subprime mortgage market became the catalyst for a global liquidity crisis. Banks began to hoard cash and refused to lend to other banks through the interbank market at anything other than extremely punitive interest rates. This caused severe difficulties for a U.K. mortgage bank, Northern Rock. Northern Rock's mortgage book had expanded rapidly in the preceding years as it borrowed aggressively from the money markets. It is now clear that this expansion was at the expense of loan quality. The U.K. regulatory authority, the Financial Services Authority (FSA), later reported in 2008 that Northern Rock's lending practices did not pay due regard to either the credit quality of the mortgagees or the values of the properties on which the mortgages were secured. Being at the worst end of banking practice and relying heavily on international capital markets for its funding, Northern Rock was therefore very susceptible to a global reduction in liquidity. As the liquidity crisis took hold, Northern Rock found that it could not replace its maturing money market borrowings. On 12 September 2007, in desperate need of liquidity, Northern Rock's board approached the U.K. central bank to ask for the necessary funds.

However, the news of Northern Rock's perilous liquidity position became known by the public and, more pertinently, by Northern Rock's retail depositors. On 14 September, having heard the news, depositors began to form lines outside Northern Rock branches as they tried to withdraw their savings. On that day, it was estimated that Northern Rock depositors withdrew around £1 billion, representing 5 percent of Northern Rock's deposits. Further panic ensued as investors in Internet-only Northern Rock accounts could not withdraw their money because of the collapse of Northern Rock's website. A further £1 billion was withdrawn over the next two days.

Northern Rock's share price dropped rapidly, as did the share prices of other similar U.K. banks. The crisis therefore threatened to engulf more than one bank. To prevent contagion, the Chancellor of the Exchequer announced on 17 September that the U.K. government would guarantee all Northern Rock deposits. This announcement was enough to stabilize the situation, and given that lending to Northern Rock was now just like lending to the government, deposits actually started to rise again.

Eventually Northern Rock was nationalized by the U.K. government, with the hope that at some time in the future it could be privatized once its balance sheet had been repaired.

EXHIBIT 7-5 Banking Supervision in the G-10

Country	Institution(s)
Belgium	Banking and Finance Commission
Canada	Office of the Superintendent of Financial Institutions
France	Commission Bancaire
Germany	Federal Banking Supervisory Office; Deutsche Bundesbank
Italy	Bank of Italy
Japan	Financial Services Agency
Netherlands	Bank of Netherlands
Sweden	Swedish Financial Supervisory Authority
Switzerland	Federal Commission
United Kingdom	Bank of England
United States	Office of the Comptroller of the Currency; Federal Reserve; Federal Deposit Insurance Corporation

Central banks are also often charged by the government to supervise the banking system, or at least to supervise those banks that they license to accept deposits. However, in some countries, this role is undertaken by a separate authority, and in still other countries, the central bank can be jointly responsible with another body for the supervision of its banks.

Exhibit 7-5 lists the banking supervisors in the G-10 countries (including Switzerland, added in 1964 as the 11th country); central banks are underlined. As the exhibit shows, most but not all banking systems have a single supervisor, which is not necessarily a central bank. A few countries, such as Germany and the United States, have more than one supervisor.

The United Kingdom is an interesting case study in this regard. Until May 1997, the Bank of England had statutory responsibility for banking supervision in the United Kingdom. In May 1997, banking supervision was removed from the Bank of England and assigned to a new agency, the Financial Services Authority (FSA). However, the removal of responsibility for banking supervision from the central bank was seen by some as being a contributory factor in the run on the mortgage bank Northern Rock, and generally as a contributory factor in the recent banking crisis. Because of this perceived weakness in the separation of the central bank from banking supervision, the United Kingdom's coalition government, which was elected in May 2010, returned the role of banking supervision to the Bank of England.

Perhaps the least appreciated role of a central bank is its role in the **payments system**. Central banks are usually asked to oversee, regulate, and set standards for a country's payments system. Every day millions of financial transactions take place in a modern economy. For the system to work properly, procedures must be robust and standardized. The central bank will usually oversee the payments system and will also be responsible for the successful introduction of any new processes. Given the international nature of finance, the central bank will also be responsible for coordinating payments systems internationally with other central banks.

Most central banks will also be responsible for managing their country's **foreign currency reserves** and also its gold reserves. With regard to the latter, even though countries abandoned the gold standard in the early part of the twentieth century, the world's central bankers still hold large quantities of gold. Therefore, if central banks were to decide to sell significant proportions of their gold reserves, that action could potentially depress gold prices.

Finally, central banks are usually responsible for the operation of a country's monetary policy. This is arguably the highest-profile role that these important organizations assume. Recall that monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy. As the monopoly supplier of a country's currency, central banks are in the ideal position to determine monetary policy and implement it.

To summarize, central banks assume a range of roles and responsibilities. They do not all assume responsibility for the supervision of the banks, but all of the other roles listed here are normally assumed by the central bank:

- Monopoly supplier of the currency.
- Banker to the government and the bankers' bank.
- Lender of last resort.
- Regulator and supervisor of the payments system.
- Conductor of monetary policy.
- Supervisor of the banking system.

2.3. Objectives of Monetary Policy

Central banks fulfill a variety of important roles, but for what overarching purpose? A brief perusal of the websites of the world's central banks will reveal a wide range of explanations of their objectives. Their objectives are clearly related to their roles, and so there is frequent mention of objectives related to the stability of the financial system and to the payments system. Some central banks are charged with doing all they can to maintain full employment and output. But some also have related but less tangible roles, like "maintaining confidence in the financial system," or even to "promote understanding of the financial sector." But there is one overarching objective that most seem to acknowledge explicitly, and that is the objective of maintaining **price stability**.

So although central banks usually have to perform many roles, most specify an overarching objective. Exhibit 7-6 lists what we might call the primary objective(s) of a number of central banks, from both developed and developing economies.

EXAMPLE 7-5 Central Banks

1. A central bank is normally *not* the:
   A. lender of last resort.
   B. banker to the government and banks.
   C. body that sets tax rates on interest on savings.

2. Which of the following *best* describes the overarching, long-run objective of most central banks?
   A. Price stability
   B. Fast economic growth
   C. Current account surplus

Solution to 1: C is correct. A central bank is normally the lender of last resort and the banker to the banks and government, but the determination of all tax rates is normally the preserve of the government and is a fiscal policy issue.

Solution to 2: A is correct. Central banks normally have a variety of objectives, but the overriding one is nearly always price stability.

EXHIBIT 7-6 The Objectives of Central Banks

The Central Bank of Brazil
Its "institutional mission" is to "ensure the stability of the currency's purchasing power and a solid and efficient financial system."

The European Central Bank
"[T]o maintain price stability is the primary objective of the Euro system and of the single monetary policy for which it is responsible. This is laid down in the Treaty on the Functioning of the European Union, Article 127 (1).

"Without prejudice to the objective of price stability," the euro system will also "support the general economic policies in the Community with a view to contributing to the achievement of the objectives of the Community." These include a "high level of employment" and "sustainable and non-inflationary growth."

The U.S. Federal Reserve
"The Federal Reserve sets the nation's monetary policy to promote the objectives of maximum employment, stable prices, and moderate long-term interest rates."

The Reserve Bank of Australia
"It is the duty of the Reserve Bank Board, within the limits of its powers, to ensure that the monetary and banking policy of the Bank is directed to the greatest advantage of the people of Australia and that the powers of the Bank . . . are exercised in such a manner as, in the opinion of the Reserve Bank Board, will best contribute to:

a. the stability of the currency of Australia;
b. the maintenance of full employment in Australia; and
c. the economic prosperity and welfare of the people of Australia."

The Bank of Korea
"The primary purpose of the Bank, as prescribed by the Bank of Korea Act of 1962, is the pursuit of price stability."

Source: Central bank websites found at www.bis.org/cbanks.htm.

As we have already discussed, one of the essential features of a monetary system is that the medium of exchange should have a relatively stable value from one period to the next. Arguably, then, the overarching goal of most central banks in maintaining price stability is the associated goal of controlling inflation. But before we explore the tools central banks use to control inflation, we should first consider the potential costs of inflation. In other words, we should ask why it is that central bankers believe that it is so important to control a nominal variable.

2.3.1. The Costs of Inflation

Huge efforts have been put into controlling inflation since the major economies experienced such high levels of inflation in the 1970s. From the early 1970s, then, inflation has been seen as a very bad thing. But why? What are the costs of inflation? The debate around the costs of inflation really centers on the distinction between expected inflation and unexpected inflation. **Expected inflation** is clearly the level of inflation that economic agents expect in the future. **Unexpected (unanticipated) inflation** can be defined as the level of inflation that we experience that is either below or above that which we expected; it is the component of inflation that is a surprise.

At a microeconomic level, high inflation means that businesses constantly have to change the advertised prices of their goods and services. These are known as **menu costs**. There also exists what economists refer to as "shoe leather" costs of inflation. In times of high inflation, people would naturally tend to hold less cash and would therefore wear out their shoe leather (or more likely the engines of their cars) in making frequent trips to the bank to withdraw cash. But these are relatively old arguments used to demonstrate that inflation is bad. In a modern economy, with the Internet and with transactions becoming increasingly cashless, these costs associated with inflation will be lower today than they may have been in the past.

To demonstrate the potentially more significant costs of inflation, consider the following. Imagine a world where inflation is high but where all prices (including asset prices) in an economy are perfectly indexed to inflation, and that technology has eliminated the issues surrounding the menu and shoe leather costs of inflation. In such a world, would economic agents care about inflation? Probably not. If the average price of goods and services rose by 10 percent, people's salaries (and all other prices) would rise by the same amount, which would therefore make economic agents indifferent to the rise in prices.

In practice, however, all prices, wages, salaries, rents, and so forth are not indexed, in which case economic agents would certainly need to think about inflation more carefully. But what if inflation in this world where prices are no longer perfectly indexed is high but perfectly predictable? In this alternative imaginary world, economic agents would have to think about inflation, but not too hard as long as they were capable of calculating the impact of the known inflation rate on all future prices. So, if everyone knew that inflation was going to be 10 percent over the next year, then everyone could bargain for a 10 percent increase in their salaries to accommodate this, and companies could plan to put up the prices of their goods and services by 10 percent. Actually, in this world, an expectation of 10 percent inflation would become a self-fulfilling prophecy.

However, economic agents would worry about inflation in a world where all prices were not indexed and, crucially, where inflation was high and unpredictable. In fact, this is a crude description of the inflationary backdrop in many developed economies over the 1970s and 1980s, including those of the United States, France, the United Kingdom, Italy, and Canada.

Arguably it is unexpected inflation that is most costly. Inflation that is fully anticipated can be factored into wage negotiations and priced into business and financial contracts. But when inflation turns out to be higher than is anticipated, then borrowers benefit at the expense of lenders because the real value of their borrowing declines. Conversely, when inflation is lower

than is anticipated, lenders benefit at the expense of borrowers because the real value of the payment on debts rises. Furthermore, if inflation is very uncertain or very volatile, then lenders will ask for a premium to compensate them for this uncertainty. As a result, the costs of borrowing will be higher than would otherwise have been the case. Higher borrowing costs could in turn reduce economic activity, for example, by discouraging investment.

It is also possible that **inflation uncertainty** can exacerbate the economic cycle. Inflation uncertainty is the degree to which economic agents view future rates of inflation as hard to forecast. Take for example the case of an imaginary television set manufacturer. Suppose one day that the manufacturer looks out at the market for television sets and sees that their market price has risen by 10 percent. Armed with this information, the manufacturer assumes that there has been an increase in demand for television sets or maybe a reduction in supply. So to take advantage of the new, higher prices, the manufacturer extends the factory, employs more workers, and begins to produce more television sets.

Having now increased the output of the factory, the manufacturer then attempts to sell the extra television sets that the factory has produced. But to their horror, the firm's managers find out that there is no extra demand for television sets. Instead, the 10 percent rise in television prices was caused by a generalized 10 percent increase in all consumer prices across the economy. The manufacturer realizes that it has surplus stock, surplus factory capacity, and too many workers. So, it cuts back on production, lays off some of the workforce, and realizes that it won't need to invest in new plant or machinery for a long time.

This example emphasizes the potentially destabilizing impact of unexpected inflation. It demonstrates how unanticipated inflation can reduce the information content of market prices for economic agents. If we scale this example up, it should not be too difficult to imagine how unanticipated increases or decreases in the general price level could help to exacerbate—and in some extreme cases cause—economic booms and busts.

Over the past two to three decades the consensus among economists has been that unanticipated and high levels of inflation can have an impact on real things like employment, investment, and profits, and therefore that controlling inflation should be one of the main goals of macroeconomic policy. In summary:

Expected inflation can give rise to:

- Menu costs.
- Shoe leather costs.

Unanticipated (unexpected) inflation can in addition:

- Lead to inequitable transfers of wealth between borrowers and lenders (including losses to savings).
- Give rise to risk premiums in borrowing rates and the prices of other assets.
- Reduce the information content of market prices.

2.3.2. Monetary Policy Tools[5]

Central banks have three primary tools available to them: open market operations, the refinancing rate, and reserve requirements.

[5]Monetary policy tools and operations often vary considerably from economy to economy. We have tried to describe the generics of the process here. For a more detailed review of monetary operations across the world, see Gray and Talbot (2006).

2.3.2.1. Open Market Operations One of the most direct ways for a central bank to increase or reduce the amount of money in circulation is via **open market operations**. Open market operations involve the purchase and sale of government bonds from and to commercial banks and designated market makers. For example, when the central bank buys government bonds from commercial banks, this increases the reserves of private-sector banks on the asset side of their balance sheets. If banks then use these surplus reserves by increasing lending to corporations and households, then via the money multiplier process explained in Section 2.1.2, broad money growth expands. Similarly, the central bank can sell government bonds to commercial banks. By doing this, the central bank causes the reserves of commercial banks to decline, reducing their capacity to make loans (i.e., create credit) to households and corporations and thus causing broad money growth to decline through the money multiplier mechanism. In using open market operations, the central bank may target a desired level of commercial bank reserves or a desired interest rate for these reserves.

2.3.2.2. The Central Bank's Policy Rate The most obvious expression of a central bank's intentions and views comes via the interest rate it sets. The name of the **official interest rate** (or **official policy rate** or just **policy rate**) varies from central bank to central bank, but its purpose is to influence short- and long-term interest rates and ultimately real economic activity.

The interest rate that a central bank sets and that it announces publicly is normally the rate at which it is willing to lend money to the commercial banks (although practices do vary from country to country). This policy rate can be achieved by using short-term collateralized lending rates, known as **repo rates**. For example, if the central bank wishes to increase the supply of money, it might buy bonds (usually government bonds) from the banks, with an agreement to sell them back at some time in the future. This transaction is known as a **repurchase (repo) agreement**. Normally, the maturity of repo agreements ranges from overnight to two weeks. In effect, this represents a secured loan to the banks, and the lender (in this case the central bank) earns the repo rate.

Suppose that a central bank announces an increase in its official interest rate. Commercial banks would normally increase their **base rates** at the same time. A commercial bank's base rate is the reference rate on which it bases lending rates to all other customers. For example, large corporate clients might pay the base rate plus 1 percent on their borrowing from a bank, while the same bank might lend money to a small corporate client at the base rate plus 3 percent. But why would commercial banks immediately increase their base or reference rates just because the central bank's refinancing rate had increased?

The answer is that commercial banks would not want to have lent at a rate of interest that would be lower than they might be charged by the central bank. Effectively, the central bank can force commercial banks to borrow from it at this rate because it can conduct open market operations that create a shortage of money, forcing the banks to sell bonds to it with a pre-agreed repurchase price (i.e., do a repurchase agreement). The repo rate would be such that the central bank would earn the official refinancing rate on the transactions.

The names of the central banks' official refinancing rates vary. The Bank of England's refinancing rate is the **two-week repo rate**. In other words, the Bank of England fixes the rate at which it is willing to lend two-week money to the banking sector. The official policy rate of the European Central Bank (ECB) is known as the **refinancing rate** and defines the rate at which it is willing to lend short-term money to the euro area banking sector.

The corresponding rate in the United States is the **discount rate**, which is the rate for member banks borrowing directly from the Federal Reserve System. But the most important interest rate used in U.S. monetary policy is the federal funds rate. The **federal funds rate** (or

fed funds rate) is the interbank lending rate on overnight borrowings of reserves. The Federal Open Market Committee (FOMC) seeks to move this rate to a target level by reducing or adding reserves to the banking system by means of open market operations. The level of the rate is reviewed by the FOMC at its meetings held every six weeks (although the target can be changed between meetings, if necessary).

Through the setting of a policy rate, a central bank can manipulate the amount of money in the money markets. Generally speaking, the higher the policy rate, the higher the potential penalty that banks will have to pay to the central bank if they run short of liquidity, the greater will be their willingness to reduce lending, and the more likely it will be that broad money growth will shrink.

2.3.2.3. Reserve Requirements The third primary way in which central banks can limit or increase the supply of money in an economy is via their **reserve requirements**. We have already seen that the money creation process is more powerful the lower the percentage reserve requirement of banks. So, a central bank could restrict money creation by raising the reserve requirements of banks. However, this policy tool is not used much nowadays in developed economies. Indeed, some central banks, such as the Bank of England, do not even set minimum reserve requirements for the banks under their jurisdiction anymore. Changing reserve requirements frequently is disruptive for banks. For example, if a central bank increased the reserve requirements, a bank that was short on reserves might have to cease its lending activities until it had built up the necessary reserves, because deposits would be unlikely to rise quickly enough for the bank to build its reserves in this way. However, reserve requirements are still actively used in many developing countries to control lending—for example in China and in India—and they remain a potential policy tool for those central banks that do not currently use it.

To summarize, central banks can manipulate the money supply in one of three ways:

1. Open market operations.
2. Its official policy rate and associated actions in the repo market.
3. Manipulation of official reserve requirements.

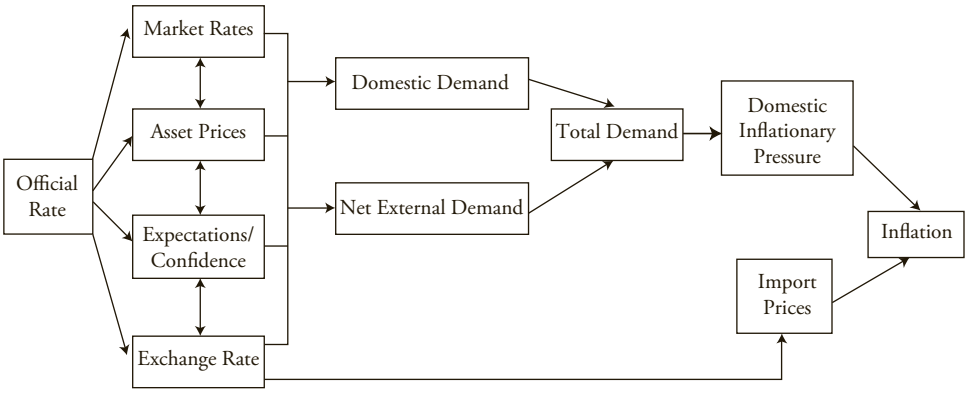
2.3.3. The Transmission Mechanism

The overarching goal of a central bank is to maintain price stability. We demonstrated earlier how a central bank can manipulate the money supply and growth of the money supply. We also indicated how policy rates set and targeted by the central banks are usually very short-term in nature; often they target overnight interest rates. However, most businesses and individuals in the real economy borrow and lend over much longer time frames than this. It may not be obvious, then, how changing short-term interest rates can influence the real economy, particularly if money neutrality holds in the long run. The fact that central bankers believe that they can affect real economic variables, in particular economic growth, by influencing broad money growth suggests that they believe that money is not neutral—at least not in the short run.

Exhibit 7-7 presents a stylized representation of the **monetary transmission mechanism**. This is the process whereby a central bank's interest rate gets transmitted through the economy and ultimately affects the rate of increase of prices—that is, inflation.

Suppose that a central bank announces an increase in its official interest rate. The implementation of the policy may begin to work through the economy via four interrelated channels. Those channels include bank lending rates, asset prices, agents' expectations, and

EXHIBIT 7-7 A Stylized Representation of the Monetary Transmission Mechanism



Source: Bank of England.

exchange rates. First, as described earlier, the base rates of commercial banks and interbank rates should rise in response to the increase in the official rate. Banks would, in turn, increase the cost of borrowing for individuals and companies over both short- and long-term horizons. Businesses and consumers would then tend to borrow less as interest rates rise. An increase in short-term interest rates could also cause the prices of such assets as bonds or the value of capital projects to fall as the discount rate for future cash flows rises.

Market participants would then come to the view that higher interest rates will lead to slower economic growth, reduced profits, and reduced borrowing to finance asset purchases. Exporters' profits might decline if the rise in interest rates causes the country's exchange rate to appreciate, because this would make domestic exports more expensive to overseas buyers and dampen demand to purchase them. The fall in asset prices as well as an increase in prices would reduce household financial wealth and therefore lead to a reduction in consumption growth. Expectations regarding interest rates can play a significant role in the economy. Often companies and individuals will make investment and purchasing decisions based on their interest rate expectations, extrapolated from recent events. If the central bank's interest rate move upward is widely expected to be followed by other interest rate increases, investors and companies will act accordingly. Consumption, borrowing, and asset prices may all decline as a result of the revision in expectations.

There is a whole range of interconnected ways in which a rise in the central bank's policy rate can reduce real domestic demand and net external demand (that is, the difference between export and import consumption). Weaker total demand would tend to put downward pressure on the rate of domestic inflation—as would a stronger currency, which would reduce the prices of imports. Taken together, these might begin to put downward pressure on the overall measure of inflation.

To summarize, the central bank's policy rate works through the economy via any one, and often all, of the following interconnected channels:

- Short-term interest rates.
- Changes in the values of key asset prices.
- The exchange rate.
- The expectations of economic agents.

EXAMPLE 7-6 Central Bank Tools

1. Which of the following variables are *most likely* to be affected by a change in a central bank's policy rate?
   A. Asset prices only.
   B. Expectations about future interest rates only.
   C. Both asset prices and expectations about future interest rates.
2. Which of the following does a central bank seek to influence directly via the setting of its official interest rate?
   A. Inflation expectations.
   B. Import prices.
   C. Domestic inflation.

Solution to 1: C is correct. The price of equities, for example, might be affected by the expectation of future policy interest rate changes. In other words, a rate change may be taken as a signal of the future stance of monetary policy—contractionary or expansionary.

Solution to 2: A is correct. By setting its official interest rate, a central bank could expect to have a direct influence on inflation expectations—as well as on other market interest rates, asset prices, and the exchange rate (where this is freely floating). If it can influence these factors, it might ultimately hope to influence import prices (via changes in the exchange rate) and also domestically generated inflation (via its impact on domestic and/or external demand). The problem is that the workings of the transmission mechanism—from the official interest rate to inflation—are complex and can change over time.

2.3.4. Inflation Targeting

Over the 1990s, a consensus began to build among both central bankers and politicians that the best way to control inflation and thereby maintain price stability was to target a certain level of inflation and to ensure that this target was met by monitoring a wide range of monetary, financial, and real economic variables. Nowadays, inflation-targeting frameworks are the cornerstone of monetary policy and macroeconomic policy in many economies. Exhibit 7-8 shows the growth in the number of inflation-targeting monetary policy regimes over time.

The inflation-targeting framework that is now commonly practiced was pioneered in New Zealand. In 1988, the New Zealand Minister of Finance, Roger Douglas, announced that economic policy would focus on bringing inflation down from the prevailing level of around 6 percent to a target range of 0 to 2 percent. This goal was given legal status by the Reserve Bank of New Zealand Act 1989. As part of the Act, the Reserve Bank of New Zealand (RBNZ) was given the role of pursuing this target. The bank was given **operational independence**; it was free to set interest rates in the way that it thought would best meet the inflation target. Although the RBNZ had independent control of monetary policy, it was still

EXHIBIT 7-8 The Progressive Adoption of Inflation Targeting by Central Banks

1989	New Zealand				
1990	Chile	Canada			
1991	Israel	United Kingdom			
1992	Sweden	Finland	Australia		
1995	Spain				
1998	Czech Republic	South Korea	Poland		
1999	Mexico	Brazil	Colombia	ECB	
2000	South Africa	Thailand			
2001	Iceland	Norway	Hungary	Peru	Philippines
2005	Guatemala	Indonesia	Romania		
2006	Turkey	Serbia			
2007	Ghana				

Note: Spain and Finland later joined the European Economic and Monetary Union. ECB refers to the European Central Bank.

Sources: For 2001 and earlier, Truman (2003). For 2002 to 2007, Roger (2010).

accountable to the government and was charged with communicating its decisions in a clear and transparent way. As Exhibit 7-8 shows, the New Zealand model was widely copied.

Although these inflation-targeting regimes vary a little from economy to economy, their success is thought to depend on three key concepts: central bank independence, credibility, and transparency.

2.3.4.1. Central Bank Independence[6] In most cases, the central bank that is charged with targeting inflation has a degree of independence from its government. This independence is thought to be important. It is conceivable that politicians could announce an inflation target and direct the central bank to set interest rates accordingly. Indeed, this was the process adopted in the United Kingdom between 1994 and 1997. But politicians have a constant eye on reelection and might be tempted, for example, to keep rates too low in the lead-up to an election in the hope that this might help their reelection prospects. As a consequence, this might lead to higher inflation. Thus, it is now widely believed that monetary policy decisions should rest in the hands of an organization that is remote from the electoral process. The central bank is the natural candidate to be the monopoly supplier of a currency.

However, there are degrees of independence. For example, the head of the central bank is nearly always chosen by government officials. The chairman of the U.S. Federal Reserve's Board of Governors is appointed by the President of the United States, the head of the ECB is chosen by the committee of euro area finance ministers, and the governor of the Bank of England is chosen by the Chancellor of the Exchequer. So, in practice, separating control from political influence completely is probably an impossible (although a desirable) goal.

[6]For information about the degree of independence of any central bank, the roles that it assumes in an economy, and the framework in which it operates, analysts should go to a central bank's website. A list of central bank websites can be found at www.bis.org/cbanks.htm.

There are further degrees of independence. Some central banks are both operationally and **target independent**. This means that they not only decide the level of interest rates, but they also determine the definition of inflation that they target, the rate of inflation that they target, and the horizon over which the target is to be achieved. The ECB has independence of this kind. By contrast, other central banks—including those in New Zealand, Sweden, and the United Kingdom—are tasked to hit a definition and level of inflation determined by the government. These central banks are therefore only operationally independent.

2.3.4.2. Credibility The independence of the central bank and public confidence in it are key in the design of an inflation-targeting regime.

To illustrate the role of credibility, suppose that instead of the central bank, the government assumes the role of targeting inflation but the government is heavily indebted. Given that higher inflation reduces the real value of debt, the government would have an incentive to avoid reaching the inflation target or to set a high inflation target such that price stability and confidence in the currency could be endangered. As a result, few would believe the government was really intent on controlling inflation; thus, the government would lack credibility. Many governments have very large levels of debt, especially since the 2008–2009 global financial crisis. In such a situation, economic agents might expect a high level of inflation, regardless of the stated target. The target might have little credibility if the organization's likelihood of sticking to it is in doubt.

If a respected central bank assumes the inflation-targeting role and if economic agents believe that the central bank will hit its target, the belief itself could become self-fulfilling. If everyone believes that the central bank will hit an inflation target of 2 percent next year, this expectation might be built into wage claims and other nominal contracts that would make it hit the 2 percent target. It is for this reason that central bankers pay a great deal of attention to inflation expectations. If these expectations were to rise rapidly, perhaps following a rapid increase in oil prices, unchecked expectations could get embedded into wage claims and eventually cause inflation to rise.

2.3.4.3. Transparency One way of establishing credibility is for a central bank to be transparent in its decision making. Many, if not all, independent inflation-targeting central banks produce a quarterly assessment of their economies. These **inflation reports**, as they are usually known, give central banks' views on the range of indicators that they watch when they come to their (usually) monthly interest rate decision. They will consider and outline their views on the following subjects, usually in this order:

- Broad money aggregates and credit conditions.
- Conditions in financial markets.
- Developments in the real economy (e.g., the labor market).
- Evolution of prices.

Consideration of all of these important components of an economy is then usually followed by a forecast of growth and inflation over a medium-term horizon, usually two years.

By explaining their views on the economy and by being transparent in decision making, the independent, inflation-targeting central banks seek to gain reputation and credibility, making it easier to influence inflation expectations and hence ultimately easier to meet the inflation target.

EXHIBIT 7-9 A Range of Inflation Targets

Country/Region	
Australia	The Australian Federal Reserve's target is inflation between 2 percent and 3 percent.
Canada	The Bank of Canada's target is CPI inflation between 1 percent and 3 percent.
Euro area	The ECB's target is CPI inflation below a ceiling of 2 percent.
South Korea	The Bank of Korea's target for 2010–2012 is CPI inflation within ±1 percentage point of 3 percent.
New Zealand	The Reserve Bank of New Zealand's target is inflation between 1 percent and 3 percent.
Sweden	The Riksbank's target is CPI inflation within ±1 percentage point of 2 percent.
United Kingdom	The Bank of England's target is CPI inflation within ±1 percentage point of 2 percent.

Source: Central bank websites (www.bis.org/cbanks.htm).

The Target Whether the target is set by the central bank or by the government for the central bank to hit, the level of the target and the horizon over which the target is to be hit are crucial considerations in all inflation-targeting frameworks.

Exhibit 7-9 shows that many central banks in developing economies target an inflation rate of 2 percent based on a consumer price index. Given that the operation of monetary policy is both art and science, the banks are normally allowed a range around the central target of +1 percent or −1 percent. For example, with a 2 percent target, they would be tasked to keep inflation between 1 percent and 3 percent. But why target 2 percent and not zero percent?

The answer is that aiming to hit zero percent could result in negative inflation, known as **deflation**. One of the limitations of monetary policy that we discuss later is its ability or inability to deal with periods of deflation. If deflation is something to be avoided, why not target 10 percent? The answer to this question is that levels of inflation that high would not be consistent with price stability; such a high inflation rate would further tend to be associated with high inflation volatility and uncertainty. Central bankers seem to agree that 2 percent is far enough away from the risks of deflation and low enough not to lead to destabilizing inflation shocks.

Finally, we should keep in mind that the headline inflation rate that is announced in most economies every month, and which is the central bank's target, is a measure of how much a basket of goods and services has risen over the previous 12 months. It is history. Furthermore, interest rate changes made today will take some time to have their full effect on the real economy as they make their way through the monetary transmission mechanism. It is for these two reasons that inflation targeters do not target current inflation but instead usually focus on inflation two years ahead.

Although inflation-targeting mandates may vary from country to country, they have common elements: the specification of an explicit inflation target, with permissible bounds, and a requirement that the central bank should be transparent in its objectives and policy actions. This is all usually laid out in legislation that imposes statutory obligations on the

central bank. As mentioned earlier, New Zealand pioneered the inflation-targeting approach to monetary policy that has since been copied widely. Exhibit 7-10 presents New Zealand's Policy Targets Agreement, which specifies the inflation-targeting mandate of its central bank, the Reserve Bank of New Zealand.

To summarize, an inflation-targeting framework normally has the following set of features:

- An independent and credible central bank.
- A commitment to transparency.
- A decision-making framework that considers a wide range of economic and financial market indicators.
- A clear, symmetric, and forward-looking medium-term inflation target, sufficiently above zero percent to avoid the risk of deflation but low enough to ensure a significant degree of price stability.

Indeed, independence, credibility, and transparency are arguably the crucial ingredients for an effective central bank, regardless of whether they target inflation.

Main Exceptions to the Inflation-Targeting Rule Although the practice of inflation targeting is widespread, there are two prominent central banks that have not adopted a formal inflation target along the lines of the New Zealand model: the Bank of Japan and the U.S. Federal Reserve System.

The Bank of Japan

Japan's central bank, the Bank of Japan (BoJ), does not target an explicit measure of inflation. Japan's government and its monetary authorities have been trying to combat deflation for much of the past two decades. However, despite their efforts—including the outright printing of money—inflation has remained very weak. Inflation targeting is seen very much as a way of combating and controlling inflation; as such, it would seem to have no place in an economy that suffers from persistent deflation.

Some economists have argued, however, that an inflation target is exactly what the Japanese economy needs. By announcing that positive inflation of, say, 3 percent is desired by the central bank, this might become a self-fulfilling prophecy if Japanese consumers and companies factor this target into nominal wage and price contracts. But for economic agents to believe that the target will be achieved, they have to believe that the central bank is capable of achieving it. Given that the BoJ has failed to engineer persistent, positive inflation, it is debatable how much credibility Japanese households and corporations would afford such an inflation-targeting policy.

The U.S. Federal Reserve System

It is perhaps rather ironic that the world's most influential central bank, the U.S. Federal Reserve, which controls the supply of the world's de facto reserve currency, the U.S. dollar, does not have an explicit inflation target. However, it is felt that the single-minded pursuit of inflation might not be compatible with the Fed's statutory goal as laid out in the Federal Reserve Act, which charges the Fed's board to "promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates."

EXHIBIT 7-10 New Zealand's Policy Targets Agreement

This agreement between the Minister of Finance and the Governor of the Reserve Bank of New Zealand (the Bank) is made under section 9 of the Reserve Bank of New Zealand Act 1989 (the Act). The Minister and the Governor agree as follows:

1. Price stability
   a. Under Section 8 of the Act the Reserve Bank is required to conduct monetary policy with the goal of maintaining a stable general level of prices.
   b. The Government's economic objective is to promote a growing, open and competitive economy as the best means of delivering permanently higher incomes and living standards for New Zealanders. Price stability plays an important part in supporting this objective.
2. Policy target
   a. In pursuing the objective of a stable general level of prices, the Bank shall monitor prices as measured by a range of price indices. The price stability target will be defined in terms of the All Groups Consumers Price Index (CPI), as published by Statistics New Zealand.
   b. For the purpose of this agreement, the policy target shall be to keep future CPI inflation outcomes between 1 percent and 3 percent on average over the medium term.
3. Inflation variations around target
   a. For a variety of reasons, the actual annual rate of CPI inflation will vary around the medium-term trend of inflation, which is the focus of the policy target. Amongst these reasons, there is a range of events whose impact would normally be temporary. Such events include, for example, shifts in the aggregate price level as a result of exceptional movements in the prices of commodities traded in world markets, changes in indirect taxes,* significant government policy changes that directly affect prices, or a natural disaster affecting a major part of the economy.
   b. When disturbances of the kind described in clause 3(a) arise, the Bank will respond consistent with meeting its medium-term target.
4. Communication, implementation and accountability
   a. On occasions when the annual rate of inflation is outside the medium-term target range, or when such occasions are projected, the Bank shall explain in Policy Statements made under section 15 of the Act why such outcomes have occurred, or are projected to occur, and what measures it has taken, or proposes to take, to ensure that inflation outcomes remain consistent with the medium-term target.
   b. In pursuing its price stability objective, the Bank shall implement monetary policy in a sustainable, consistent and transparent manner and shall seek to avoid unnecessary instability in output, interest rates and the exchange rate.
   c. The Bank shall be fully accountable for its judgments and actions in implementing monetary policy.

*"Indirect taxes" refer to such taxes as sales taxes and value-added taxes that are levied on goods and services rather than directly on individuals and companies.

Source: www.rbnz.govt.nz/.

In other words, it has been argued that inflation targeting might compromise the goal of "maximum employment." In practice, however, the Fed has indicated that it sees core inflation measured by the personal consumption expenditure (PCE) deflator of about, or just below, 2 percent as being compatible with "stable prices." Financial markets therefore watch this U.S. inflation gauge very carefully in order to try to anticipate the rate actions of the Fed.

2.3.4.4. Monetary Policy in Developing Countries Developing economies often face significant impediments to the successful operation of any monetary policy—that is, the achievement of price stability. These include:

- The absence of a sufficiently liquid government bond market and developed interbank market through which monetary policy can be conducted.
- A rapidly changing economy, making it difficult to understand what the neutral rate might be and what the equilibrium relationship between monetary aggregates and the real economy might be.
- Rapid financial innovation that frequently changes the definition of the money supply.
- A poor track record in controlling inflation in the past, making monetary policy intentions less credible.
- An unwillingness of governments to grant genuine independence to the central bank.

Taken together, any or all of these impediments might call into question the effectiveness of any developing economy's monetary policy framework, making any related monetary policy goals difficult to achieve.

EXAMPLE 7-7 Central Bank Effectiveness

1. The reason some inflation-targeting banks may target low inflation and not zero percent inflation is *best* described by which of the following statements?
   A. Some inflation is viewed as being good for an economy.
   B. Targeting zero percent inflation runs a higher risk of a deflationary outcome.
   C. It is very difficult to eliminate all inflation from a modern economy.
2. The degree of credibility that a central bank is afforded by economic agents is important because:
   A. they are the lender of last resort.
   B. their targets can become self-fulfilling prophecies.
   C. they are the monopolistic suppliers of the currency.

Solution to 1: B is correct. Inflation targeting is art, not science. Sometimes inflation will be above target and sometimes below. Were central banks to target zero percent, then inflation would almost certainly be negative on some occasions. If a deflationary mindset then sets in among economic agents, it might be difficult for the central bank to respond to this, because it cannot cut interest rates below zero.

Solution to 2: B is correct. If a central bank operates within an inflation-targeting regime and if economic agents believe that it will achieve its target, this expectation will become embedded into wage negotiations, for example, and become a self-fulfilling prophecy. Also, banks need to be confident that the central bank will lend them money when all other sources are closed to them; otherwise, they might curtail their lending drastically, leading to a commensurate reduction in money and economic activity.

2.3.5. Exchange Rate Targeting

Many developing economies choose to operate monetary policy by targeting their currency's exchange rate, rather than an explicit level of domestic inflation. Such targeting involves setting a fixed level or band of values for the exchange rate against a major currency, with the central bank supporting the target by buying and selling the national currency in foreign exchange markets. There are recent examples of developed economies using such an approach. In the 1980s, following the failure of its policy of trying to control U.K. inflation by setting medium-term goals for money supply growth (see box earlier in this chapter, "Mrs. Thatcher's Monetary Experiment"), the U.K. government decided to operate monetary policy such that the sterling's exchange rate equaled a predetermined value in terms of German deutsche marks. The basic idea is that by tying a domestic economy's currency to that of an economy with a good track record on inflation, the domestic economy would effectively import the inflation experience of the low-inflation economy.

Suppose that a developing country wished to maintain the value of its currency against the U.S. dollar. The government and/or central bank would announce the currency exchange rate that they wished to target. To simplify matters, let us assume that the domestic inflation rates are very similar in both countries and that the monetary authorities of the developing economy have set an exchange rate target that is consistent with relative price levels in the two economies. Under these (admittedly unlikely) circumstances, in the absence of shocks, there would be no reason for the exchange rate to deviate significantly from this target level. So as long as domestic inflation closely mirrors U.S. inflation, the exchange rate should remain close to its target (or within a target band). It is in this sense that a successful exchange rate policy imports the inflation of the foreign economy.

Now suppose that economic activity in the developing economy starts to rise rapidly and that domestic inflation in the developing economy rises above the level in the United States. With a freely floating exchange rate regime, the currency of the developing economy would start to fall against the dollar. To arrest this fall and to protect the exchange rate target, the developing economy's monetary authority sells foreign currency reserves and buys its own currency. This has the effect of reducing the domestic money supply and increasing short-term interest rates. The developing economy experiences a monetary policy tightening, which, if expected to bring down inflation, will cause its exchange rate to rise against the dollar.

By contrast, in a scenario in which inflation in the developing country falls relative to inflation in the United States, the central bank would need to sell the domestic currency to support the target, tending to increase the domestic money supply and reduce the rate of interest.

In practice, the interventions of the developing economy central bank will simply stabilize the value of its currency, with many frequent adjustments. But this simplistic example should demonstrate one very important fact: *When the central bank or monetary authority chooses to*

EXHIBIT 7-11 Some Markets that Peg Currencies to the U.S. Dollar, as of December 2009

• The Netherlands Antilles	• Hong Kong SAR
• Jordan	• Lebanon
• Barbados	• Saudi Arabia
• Maldives	• Oman
• Belize	• Qatar
• The Bahamas	

target an exchange rate, interest rates and conditions in the domestic economy must adapt to accommodate this target, and domestic interest rates and money supply can become more volatile.

The monetary authority's commitment to and ability to support the exchange rate target must be credible for exchange rate targeting to be successful. If that is not the case, then speculators may trade against the monetary authority. Speculative attacks forced sterling out of the European Exchange Rate Mechanism in 1992. The fixed exchange rate regime was abandoned and the United Kingdom allowed its currency to float freely. Eventually, the U.K. government adopted a formal inflation target in 1997. Similarly, in the Asian financial crisis of 1997–1998, Thailand's central bank tried to defend the Thai baht against speculative attacks for much of the first half of 1997 but then revealed at the beginning of July that it had no reserves left. The subsequent devaluation triggered a debt crisis for banks and companies that had borrowed in foreign currency, and contagion spread throughout Asia.

Despite these risks, many economies fix their exchange rate to other currencies, most notably the U.S. dollar. Exhibit 7-11 shows a list of some of the currencies that were fixed against the U.S. dollar at the end of 2009. Other countries operate a so-called managed exchange rate policy, where they try to limit the movement of their currency by intervening in the market.

EXAMPLE 7-8 Exchange Rate Targeting

1. When the central bank chooses to target a specific value for its exchange rate:
   A. it must also target domestic inflation.
   B. it must also set targets for broad money growth.
   C. conditions in the domestic economy must adapt to accommodate this target.
2. With regard to monetary policy, what is the hoped-for benefit of adopting an exchange rate target?
   A. Freedom to pursue redistributive fiscal policy
   B. Freedom to set interest rates according to domestic conditions
   C. To import the inflation experience of the economy whose currency is being targeted

3. Which of the following is *least likely* to be an impediment to the successful implementation of monetary policy in developing economies?
   A. Fiscal deficits
   B. Rapid financial innovation
   C. Absence of a liquid government bond market

Solution to 1: C is correct. The adoption of an exchange rate target requires that the central bank set interest rates to achieve this target. If the target comes under pressure, domestic interest rates may have to rise, regardless of domestic conditions. It may have a target level of inflation in mind as well as targets for broad money growth, but as long as it targets the exchange rate, domestic inflation and broad money trends must simply be allowed to evolve.

Solution to 2: C is correct. Note that interest rates have to be set to achieve this target and are therefore subordinate to the exchange rate target and partially dependent on economic conditions in the foreign economy.

Solution to 3: A is correct. Note that the absence of a liquid government bond market through which a central bank can enact open market operations and repo transactions will inhibit the implementation of monetary policy—as would rapid financial innovation because such innovation can change the relationship between money and economic activity. Fiscal deficits, in contrast, are not normally an impediment to the implementation of monetary policy, although they could be if they were perceived to be unsustainable.

2.4. Contractionary and Expansionary Monetary Policies and the Neutral Rate

Most central banks will adjust liquidity conditions by adjusting their official policy rate.[7] When they believe that economic activity is likely to lead to an increase in inflation, they might increase interest rates, thereby reducing liquidity. In these cases, market analysts describe such actions as **contractionary** because the policy is designed to cause the rate of growth of the money supply and the real economy to contract (see Exhibit 7-7 for the possible transmission mechanism here). Conversely, when the economy is slowing and inflation and monetary trends are weakening, central banks may increase liquidity by cutting their target rate. In these circumstances, monetary policy is said to be **expansionary**.

Thus, when policy rates are high, monetary policy may be described as contractionary; when low, they may be described as expansionary. But what are they high and low in comparison to?

The **neutral rate of interest** is often taken as the point of comparison. One way of characterizing the neutral rate is to say that it is that rate of interest that neither spurs on nor

[7]Although if they have reduced their policy rate to zero percent, to increase liquidity further they have to resort to less conventional monetary policy measures.

slows down the underlying economy. As such, when policy rates are above the neutral rate, monetary policy is contractionary; when they are below the neutral rate, monetary policy is expansionary. The neutral rate should correspond to the average policy rate over a business cycle.

However, economists' views of the neutral rate for any given economy might differ, and therefore their view of whether monetary policy is contractionary, neutral, or expansionary might differ, too. What economists do agree on is that the neutral policy rate for any economy has two components:

1. Real trend rate of growth of the underlying economy.
2. Long-run expected inflation.

The real trend rate of growth of an economy is also difficult to discern, but it corresponds to that rate of economic growth that is achievable in the long run that gives rise to stable inflation. If we are thinking about an economy with a credible inflation-targeting regime, where the inflation target is, say, 2 percent per year and where an analyst believes that the economy can grow sustainably over the long term at a rate of 2.5 percent per year, then they might also estimate the neutral rate to be:

$$\text{Neutral rate} = \text{Trend growth} + \text{Inflation target} = 2\% + 2.5\% = 4.5\% \qquad (7\text{-}4)$$

The analyst would therefore describe the central bank's monetary policy as being contractionary when its policy rate is above 4.5 percent and expansionary when it is below this level.

In practice, central banks often indicate what they believe to be the neutral rate of interest for their economy, too. But determining this neutral rate is more art than science. For example, many analysts have recently revised down their estimates of trend growth for many Western countries following the collapse of the credit bubble, because in many cases the governments and private individuals of these economies are now being forced to reduce consumption levels and pay down their debts.

2.4.1. What Is the Source of the Shock to the Inflation Rate?

An important aspect of monetary policy for those charged with its conduct is the determination of the source of any shock to the inflation rate. Suppose that the monetary authority sees that inflation is rising beyond its target, or simply in a way that threatens price stability. If this rise was caused by an increase in the confidence of consumers and business leaders, which in turn has led to increases in consumption and investment growth rates, then we could think of it as being a **demand shock**. In this instance, it might be appropriate to tighten monetary policy in order to bring the inflationary pressures generated by these domestic demand pressures under control.

However, suppose instead that the rise in inflation was caused by a rise in the price of oil (for the sake of argument). In this case, the economy is facing a **supply shock**, and raising interest rates might make a bad situation worse. Consumers are already facing an increase in the cost of fuel prices that might cause profits and consumption to fall and eventually unemployment to rise. Putting up interest rates in this instance might simply exacerbate the oil price–induced downturn, which might ultimately cause inflation to fall sharply.

It is important, then, for the monetary authority to try to identify the source of the shock before engineering a contractionary or expansionary monetary policy phase.

2.5. Limitations of Monetary Policy

The limitations of monetary policy include problems in the transmission mechanism and the relative ineffectiveness of interest rate adjustment as a policy tool in deflationary environments.

2.5.1. Problems in the Monetary Transmission Mechanism

In Exhibit 7-7, we presented a stylized representation of the monetary policy transmission mechanism, including the channels of bank lending rates, asset prices, expectations, and exchange rates. The implication of the diagram is that there are channels through which the actions of the central bank or monetary authority are transmitted to both the nominal and the real economy. However, there may be some occasions when the will of the monetary authority is not transmitted seamlessly through the economy.

Suppose that a central bank raises interest rates because it is concerned about the strength of underlying inflationary pressures. Long-term interest rates are influenced by the path of expected short-term interest rates, so the outcome of the rate hike will depend on market expectations. Suppose that bond market participants think that short-term rates are already too high, that the monetary authorities are risking a recession, and that the central bank will likely undershoot its inflation target. This fall in inflation expectations could cause long-term interest rates to fall. That would make long-term borrowing cheaper for companies and households, which could in turn stimulate economic activity rather than cause it to contract.

Arguably, the more credible the monetary authority, the more stable the long end of the yield curve; moreover, the monetary authority will be more confident that its policy message will be transmitted throughout the economy. A term recently used in the marketplace is **bond market vigilantes**. These "vigilantes" are bond market participants who might reduce their demand for long-term bonds, thus pushing up their yields, if they believe that the monetary authority is losing its grip on inflation. That yield increase could act as a brake on any loose monetary policy stance. Conversely, the vigilantes may push long-term rates down by increasing their demand for long-dated government bonds if they expect that tight monetary policy is likely to cause a sharp slowdown in the economy, thereby loosening monetary conditions for long-term borrowers in the economy.

A credible monetary policy framework and authority will tend not to require the vigilantes to do the work for it.

In very extreme instances, there may be occasions where the demand for money becomes infinitely elastic—that is, where the demand curve is horizontal and individuals are willing to hold additional money balances without any change in the interest rate—so that further injections of money into the economy will not serve to further lower interest rates or affect real activity. This is known as a **liquidity trap**. In this extreme circumstance, monetary policy can become completely ineffective. The economic conditions for a liquidity trap are associated with the phenomenon of deflation.

2.5.2. Interest Rate Adjustment in a Deflationary Environment and Quantitative Easing as a Response

Deflation is a pervasive and persistent fall in a general price index and is more difficult for conventional monetary policy to deal with than inflation. This is because once the monetary authority has cut nominal interest rates to zero to stimulate the economy, it cannot cut them any further. It is at this point that the economic conditions for a liquidity trap arise.

Deflation raises the real value of debt, while the persistent fall in prices can encourage consumers to put off consumption today, leading to a fall in demand that leads to further

deflationary pressure. Thus a deflationary trap can develop, which is characterized by weak consumption growth, falling prices, and increases in real debt levels. Japan eventually found itself in such a position following the collapse of its property bubble in the early 1990s.

If conventional monetary policy—the adjustment of short-term interest rates—is no longer capable of stimulating the economy once the zero nominal interest rate bound has been reached, is monetary policy useless?

In the aftermath of the collapse of the high-tech bubble in November 2002, Federal Reserve Governor (now Chairman) Ben Bernanke gave a speech entitled "Deflation: Making Sure 'It' Doesn't Happen Here." In this speech, Bernanke stated that inflation was always and everywhere a monetary phenomenon, and he expressed great confidence that by expanding the money supply by various means (including dropping it out of a helicopter on the population below), the Federal Reserve as the monopoly supplier of money could always engineer positive inflation in the U.S. economy. He said:

> I am confident that the Fed would take whatever means necessary to prevent significant deflation in the United States and, moreover, that the U.S. central bank, in cooperation with other parts of the government as needed, has sufficient policy instruments to ensure that any deflation that might occur would be both mild and brief.

Following the collapse of the credit bubble in 2008, a number of governments along with their central banks cut rates to (near) zero, including those in the United States and the United Kingdom. However, there was concern that the underlying economies might not respond to this drastic monetary medicine, mainly because the related banking crisis had caused banks to reduce their lending drastically. In order to kick-start the process, both the Federal Reserve and the Bank of England effectively printed money and pumped it into their respective economies. This unconventional approach to monetary policy, known as **quantitative easing** (QE), is operationally similar to open market purchase operations but conducted on a much larger scale.

The additional reserves created by central banks in a policy of quantitative easing can be used to buy any assets. The Bank of England chose to buy **gilts** (bonds issued by the U.K. government), where the focus was on gilts with three to five years' maturity. The idea was that this additional reserve would kick-start lending, causing broad money growth to expand, which would eventually lead to an increase in real economic activity. But there is no guarantee that banks will respond in this way. In a difficult economic climate, it may be better to hold excess reserves rather than to lend to households and businesses that may default.

In the United States, the formal plan for QE mainly involved the purchase of mortgage bonds issued or guaranteed by Freddie Mac and Fannie Mae. Part of the intention was to push down mortgage rates to support the U.S. housing market, as well as to increase the growth rate of broad money. Before implementing this formal program, the Federal Reserve intervened in several other markets that were failing for lack of liquidity, including interbank markets and the commercial paper market. These interventions had a similar effect on the Federal Reserve's balance sheet and the money supply as the later QE program.

This first round of QE by the Federal Reserve was then followed by a further round of QE, known as QE2. In November 2010, the Federal Reserve judged that the U.S. economy had not responded sufficiently to the first round of QE (QE1). The Fed announced that it would create $600 billion and use this money to purchase long-dated U.S. Treasuries in equal tranches over the following eight months. The purpose of QE2 was to ensure that long bond

yields remained low in order to encourage businesses and households to borrow for investment and consumption purposes, respectively.

As long as they have the appropriate authority from the government, central banks can purchase any assets in a quantitative easing program. But the risks involved in purchasing assets with credit risk should be clear. In the end, the central bank is just a special bank. If it accumulates bad assets that then turn out to create losses, it could face a fatal loss of confidence in its main product: fiat money.

2.5.3. Limitations of Monetary Policy: Summary

The ultimate problem for monetary authorities as they try to manipulate the supply of money in order to influence the real economy is that they cannot control the amount of money that households and corporations put in banks on deposit, nor can they easily control the willingness of banks to create money by expanding credit. Taken together, this also means that they cannot always control the money supply. Therefore, there are definite limits to the power of monetary policy.

The Limits of Monetary Policy: The Case of Japan

The Background

Between the 1950s and 1980s, Japan's economy achieved faster real growth than any other G-7 economy. But the terrific success of the economy sowed the seeds of the problems that were to follow. The very high real growth rates achieved by Japan over four decades became built into asset prices, particularly equity and commercial property prices. Toward the end of the 1980s, asset prices rose to even higher levels when the Bank of Japan (BoJ) followed a very easy monetary policy as it tried to prevent the Japanese yen from appreciating too much against the U.S. dollar. However, when interest rates went up in 1989 and 1990 and the economy slowed, investors eventually came to believe that the growth assumptions that were built into asset prices and other aspects of the Japanese economy were unrealistic. This realization caused Japanese asset prices to collapse. For example, the Nikkei 225 stock market index reached 38,915 in 1989; by the end of March 2003, it had fallen by 80 percent to 7,972. The collapse in asset prices caused wealth to decline dramatically. Consumer confidence understandably fell sharply, too, and consumption growth slowed. Corporate spending also fell, while bank lending contracted sharply in the weak economic climate. Although many of these phenomena are apparent in all recessions, the situation was made worse when deflation set in. In an environment when prices are falling, consumers may put off today's discretionary spending until tomorrow; by doing this, however, they exacerbate the deflationary environment. Deflation also raises the real value of debts; as deflation takes hold, borrowers find the real value of their debts rising and may try to increase their savings accordingly. Once again, such actions exacerbate the recessionary conditions.

The Monetary Policy Response

Faced with such a downturn, the conventional monetary policy response is to cut interest rates to try to stimulate real economic activity. The Japanese central bank, the

Bank of Japan, cut rates from 8 percent in 1990 to 1 percent by 1996. By February 2001, the Japanese policy rate was cut to zero. Once this point is reached, a central bank cannot lower rates any further because nominal interest rates cannot be negative. The Bank of Japan has kept rates at zero since February 2001.

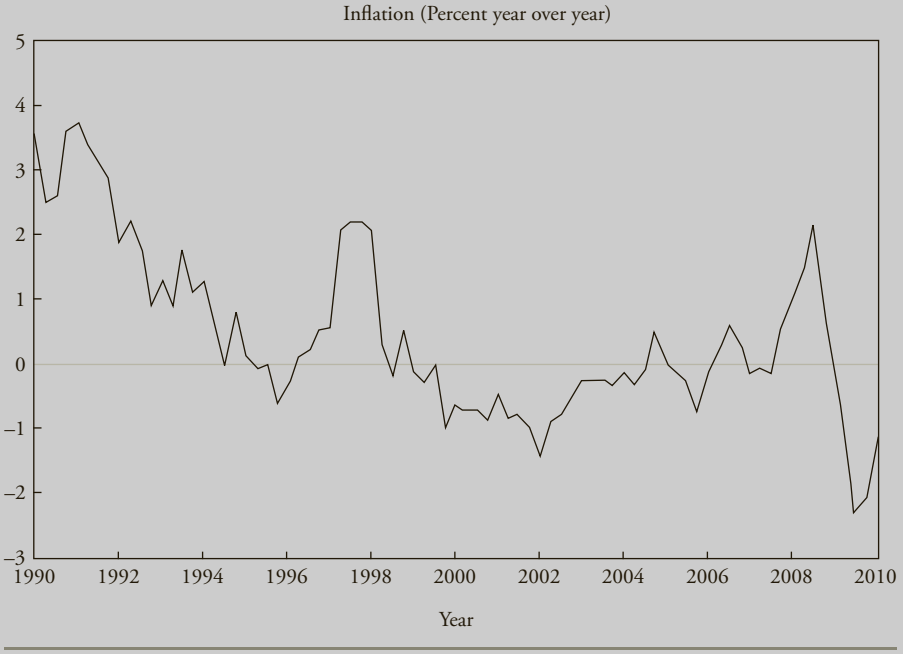
Once rates are at zero, there are two broad approaches suggested by theory, though the two are usually complementary. First, the central bank can try to convince markets that interest rates will remain low for a long time, even after the economy and inflation pick up. This will tend to lower interest rates along the yield curve. Second, the central bank can try to increase the money supply by purchasing assets from the private sector, so-called quantitative easing. The Bank of Japan did both in 2001. It embarked on a program of quantitative easing supplemented by an explicit promise not to raise short-term interest rates until deflation had given way to inflation.

Quantitative easing simply involves the printing of money by the central bank. In practice, this involved the BoJ using open market operations to add reserves to the banking system through the direct purchase of government securities in the open market.

The reserve levels became the new target. The BoJ's monetary policy committee determined the level of reserves and the quantity of bond purchases that should be undertaken, rather than voting on the policy rate.

The success of this policy is difficult to judge. As the chart in Exhibit 7-12 shows, although deflation turned to inflation for a while, it returned to deflation in 2008 and 2009 when the Japanese economy suffered a sharp recession along with much of the rest of the world. At that time, having reversed its QE policy during 2004–2008 by reducing its bond holdings, the Bank of Japan began to buy again.

EXHIBIT 7-12 Inflation and Deflation in Japan



Source: Thomson Financial.

Economists debate the point, but arguably the Bank of Japan needed to implement a much larger program of QE to eliminate deflation. Japan's program amounted to a cumulative 7 to 8 percent of GDP spread over three years, whereas the United States and United Kingdom implemented programs totaling 12 percent and 14 percent, respectively, in about 12 months during 2009 and 2010. The Japanese experience underlines the limits to the power of monetary policy.

EXAMPLE 7-9 Evaluating Monetary Policy

1. If an economy's trend GDP growth rate is 3 percent and its central bank has a 2 percent inflation target, which policy rate is *most* consistent with an expansionary monetary policy?
   A. 4 percent
   B. 5 percent
   C. 6 percent
2. An increase in a central bank's policy rate might be expected to reduce inflationary pressures by:
   A. reducing consumer demand.
   B. reducing the foreign exchange value of the currency.
   C. driving up asset prices, leading to an increase in personal sector wealth.
3. Which of the following statements *best* describes a fundamental limitation of monetary policy? Monetary policy is limited because central bankers:
   A. cannot control the inflation rate perfectly.
   B. are appointed by politicians and are therefore never truly independent.
   C. cannot control the amount of money that economic agents put in banks, nor the willingness of banks to make loans.

Solution to 1: A is correct. The neutral rate of interest, which in this example is 5 percent, is considered to be that rate of interest that neither spurs on nor slows down the underlying economy. As such, when policy rates are above the neutral rate, monetary policy is contractionary; when they are below the neutral rate, monetary policy is expansionary. It comprises two components: the real trend rate of growth of the underlying economy (in this example, 3 percent) and long-run expected inflation (in this example, 2 percent).

Solution to 2: A is correct. If an increase in the central bank's policy rate is successfully transmitted via the money markets to other parts of the financial sector, consumer demand might decline as the rate of interest on mortgages and other credit rises. This decline in consumer demand should, all other things being equal and among other effects, lead to a reduction in upward pressure on consumer prices.

Solution to 3: C is correct. Central bankers do not control the decisions of individuals and banks that can influence the money creation process.

3. FISCAL POLICY

The second set of tools used for influencing economic activity consists of the tools associated with fiscal policy. These involve the use of government spending and changing tax revenue to affect a number of aspects of the economy:

- Overall level of aggregate demand in an economy and hence the level of economic activity.
- Distribution of income and wealth among different segments of the population.
- Allocation of resources among different sectors and economic agents.

Often, a discussion of fiscal policy focuses on the impact of changes in the difference between government spending and revenue on the aggregate economy, rather than on the actual levels of spending and revenue themselves.

3.1. Roles and Objectives of Fiscal Policy

A primary aim for fiscal policy is to help manage the economy through its influence on aggregate national output, that is, real GDP.

3.1.1. Fiscal Policy and Aggregate Demand

Aggregate demand is the amount companies and households plan to spend. We can consider a number of ways that fiscal policy can influence aggregate demand. For example, an expansionary policy could take one or more of the following forms:

- Cuts in personal income tax raise disposable income with the objective of boosting aggregate demand.
- Cuts in sales (indirect) taxes to lower prices, which raises real incomes with the objective of raising consumer demand.
- Cuts in corporation (company) taxes to boost business profits, which may raise capital spending.
- Cuts in tax rates on personal savings to raise disposable income for those with savings, with the objective of raising consumer demand.
- New public spending on social goods and infrastructure, such as hospitals and schools, boosting personal incomes with the objective of raising aggregate demand.

We must stress, however, that the reliability and magnitude of these relationships will vary over time and from country to country. For example, in a recession with rising unemployment, it is not always the case that cuts in income taxes will raise consumer spending, because consumers may wish to raise their precautionary (rainy day) saving in anticipation of further deterioration in the economy. Indeed, in very general terms economists are often divided into two camps regarding the workings of fiscal policy: **Keynesians** believe that fiscal policy can have powerful effects on aggregate demand, output, and employment when there is substantial spare capacity in an economy. Monetarists believe that fiscal changes have only a temporary effect on aggregate demand and that monetary policy is a more effective tool for restraining or boosting inflationary pressures. Monetarists tend not to advocate using

monetary policy for countercyclical adjustment of aggregate demand. This intellectual division will naturally be reflected in economists' divergent views on the efficacy of the large fiscal expansions observed in many countries following the credit crisis of 2008, along with differing views on the possible impact of quantitative easing.

3.1.2. Government Receipts and Expenditure in Major Economies

In Exhibit 7-13, we present the total government revenues as a percentage of GDP for some major economies. This is the share of a country's output that is gathered by the government through taxes and such related items as fees, charges, fines, and capital transfers. It is often considered as a summary measure of the extent to which a government is involved both directly and indirectly in the economic activity of a country.

Taxes are formally defined as compulsory, unrequited payments to the general government (they are unrequited in the sense that benefits provided by a government to taxpayers are usually not related to payments). Exhibit 7-13 contains taxes on incomes and profits, social security contributions, indirect taxes on goods and services, employment taxes, and taxes on the ownership and transfer of property.

Taxes on income and profits have been fairly constant for the Organization for Economic Cooperation and Development (OECD) countries overall at around 12.5 to 13 percent of GDP since the mid-1990s, while taxes on goods and services have been steady at about 11 percent of GDP for that period. Variations between countries can be substantial; taxes on goods and services are around 5 percent of GDP for the United States and Japan but over 16 percent for Denmark.

Exhibit 7-14 shows the percentage of GDP represented by government expenditure in a variety of major economies over time. Generally, these have been fairly constant since 1995, though Germany had a particularly high number at the start of the period because of reunification costs.

Clearly, the possibility that fiscal policy can influence output means that it may be an important tool for **economic stabilization**. In a recession, governments can raise spending (**expansionary fiscal policy**) in an attempt to raise employment and output. In boom times—

EXHIBIT 7-13 General Government Revenues as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	34.5	36.1	36.5	36.4	36.0	35.3
Germany	45.1	46.4	43.6	43.7	43.8	43.8
Japan	31.2	31.4	31.7	34.5	33.5	34.4
United Kingdom	38.2	40.3	40.8	41.4	41.4	42.2
United States	33.8	35.4	33.0	33.8	34.0	32.3
OECD	37.9	39.0	37.7	38.6	38.6	37.9

Source: Organization for Economic Cooperation and Development (OECD).

EXHIBIT 7-14 General Government Expenditures as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	38.2	35.2	34.8	34.5	34.2	34.3
Germany	54.8	45.1	46.9	45.3	43.6	43.8
Japan	36.0	39.0	38.4	36.2	36.0	37.1
United Kingdom	44.1	36.6	44.0	44.1	44.2	47.5
United States	37.1	33.9	36.2	36.0	36.8	38.8
OECD	42.7	38.7	40.5	39.9	39.9	41.4

Source: Organization for Economic Cooperation and Development (OECD).

EXHIBIT 7-15 General Government Net Borrowing or Lending as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	–3.7	0.9	1.7	1.9	1.8	1.0
Germany	–9.7	1.3	–3.3	–1.6	0.2	0.0
Japan	–4.7	–7.6	–6.7	–1.6	–2.5	–2.7
United Kingdom	–5.8	3.7	–3.3	–2.7	–2.7	–5.3
United States	–3.3	1.5	–3.3	–2.2	–2.8	–6.5
OECD	–4.8	0.2	–2.7	–1.3	–1.3	–1.3

Source: Organization for Economic Cooperation and Development (OECD).

when an economy has full employment and wages and prices are rising too fast—then government spending may be reduced and taxes raised (**contractionary fiscal policy**).

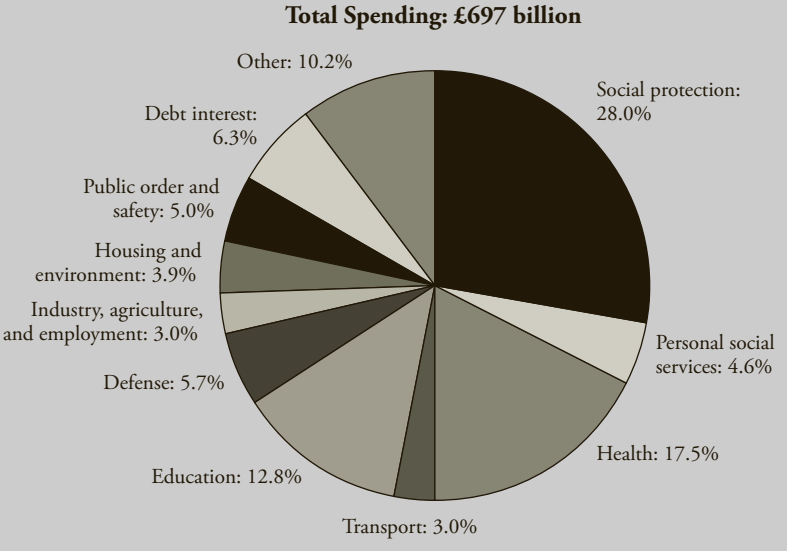
Hence, a key concept is the **budget surplus/deficit**, which is the difference between government revenue and expenditure for a fixed period of time, such as a fiscal or calendar year. Government revenue includes tax revenues net of transfer payments; government spending includes interest payments on the government debt. Analysts often focus on changes in the budget surplus or deficit from year to year as indicators of whether the fiscal policy is getting tighter or looser. An increase in a budget surplus would be associated with a contractionary fiscal policy, whereas a rise in a deficit is an expansionary fiscal policy. Of course, over the course of a business cycle the budget surplus will vary automatically in a countercyclical way. For example, as an economy slows and unemployment rises, government spending on social insurance and unemployment benefits will also rise and add to aggregate demand. This is known as an **automatic stabilizer**. Similarly, if boom conditions ensue and employment and incomes are high, then progressive income and profit taxes are rising and also act as automatic stabilizers that would reduce the growing budget surplus. The great advantage of automatic stabilizers is that they are indeed automatic, not

requiring the identification of shocks to which policy makers must consider a response. By reducing the responsiveness of the economy to shocks, these automatic stabilizers reduce output fluctuations. Automatic stabilizers should be distinguished from discretionary fiscal policies, such as changes in government spending or tax rates, which are actively used to stabilize aggregate demand. If government spending and revenues are equal, then the budget is **balanced**.

Sources and Uses of Government Cash Flows: The Case of the United Kingdom

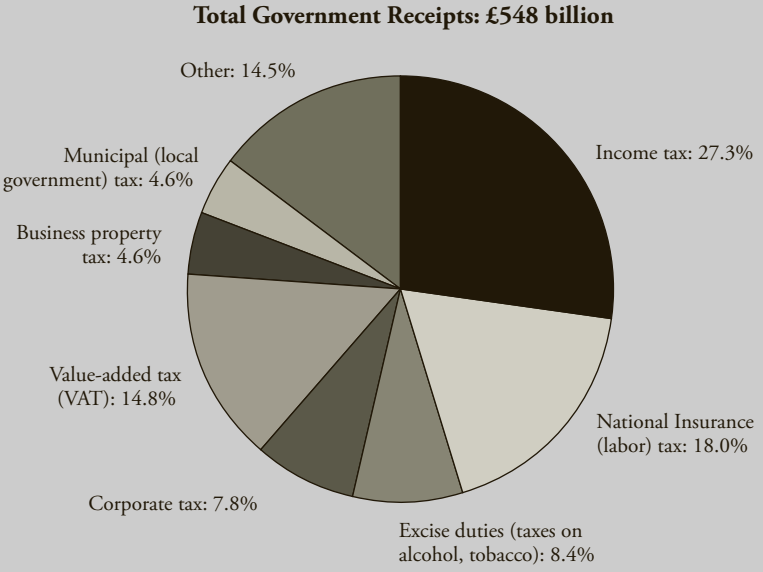
The precise components of revenue and expenditure will of course vary over time and among countries. But, as an example of the breakdown of expenditure and revenue, in Exhibits 7-16 and 7-17 we present the budget projections of the United Kingdom for 2010–2011. The budget projected that total spending would come to £697 billion, while total revenue would be only £548 billion. The government was therefore forecasting a budget shortfall of £149 billion for the fiscal year, meaning that it had an associated need to borrow £149 billion from the private sector in the United Kingdom or the private and public sectors of other economies.

EXHIBIT 7-16 Where Does the Money Go? The United Kingdom, 2010–2011



Note: "Other" includes recreation, culture, religion, public sector pensions, and general public services.
Source: HM Treasury, United Kingdom.

EXHIBIT 7-17 Where Does the Money Come From? The United Kingdom, 2010–2011



Note: "Other" includes capital taxes, stamp duties on share and property transactions, and vehicle excise duties (road tax).

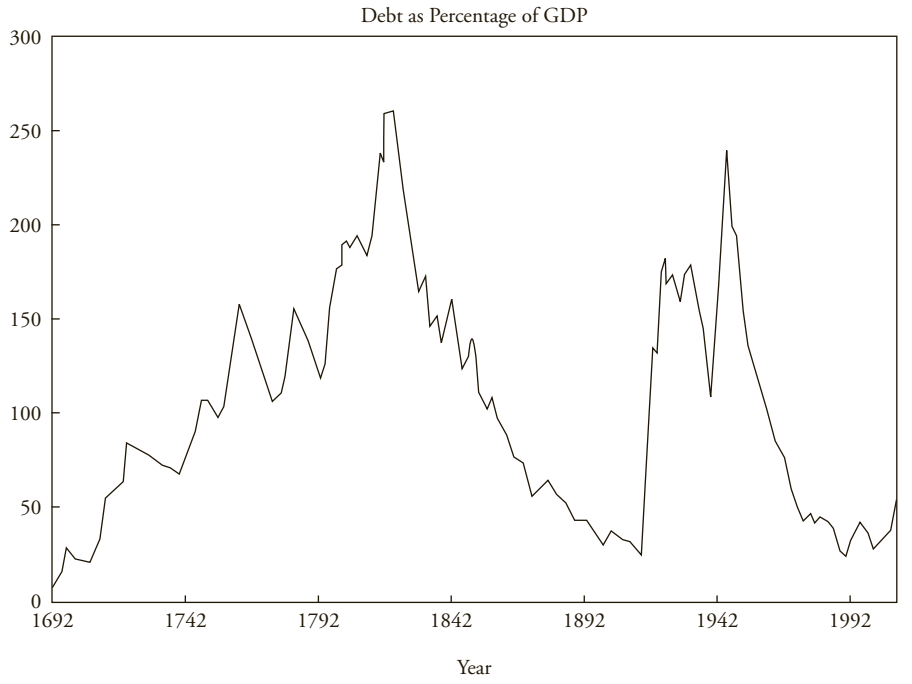
3.1.3. Deficits and the National Debt

Government deficits are the difference between government revenues and expenditures over a period of calendar time, usually a year. Government (or national) debt is the accumulation over time of these deficits. Government deficits are financed by borrowing from the private sector, often via private pension and insurance fund portfolio investments. We saw earlier that governments are more likely to have deficits than surpluses over long periods of time. As a result, there may exist a large stock of outstanding government debt owned by the private sector. This will vary as the business cycle ebbs and flows. Exhibit 7-18 shows the time path of the ratio of public debt to GDP for the United Kingdom over several hundred years. It can be clearly seen that the major cause of fluctuations in that ratio through history has been the financing of wars, in particular the Napoleonic Wars of 1799–1815 and the First and Second World Wars of 1914–1918 and 1939–1945.

With the onset of the credit crisis of 2008, governments actively sought to stimulate their economies through increased expenditures without raising taxes and revenues. This led to increased borrowing, shown in Exhibits 7-15 and 7-19, which has become a concern in the financial markets in 2010 for such countries as Greece. Indeed, between 2008 and 2009, central government debt rose from $1.2 trillion to $1.6 trillion in the United Kingdom and from $5.8 trillion to $7.5 trillion for the United States.[8]

[8] *Source:* www.oecd.org.

EXHIBIT 7-18 U.K. National Debt as Percentage of GDP, 1692–2010



Source: http://ukpublicspending.co.uk.

EXHIBIT 7-19 General Government Gross Financial Liabilities as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	42.5	25.4	16.9	16.2	15.3	14.3
Germany	55.7	60.4	71.1	69.2	65.3	68.8
Japan	86.2	135.4	175.3	172.1	167.1	172.1
United Kingdom	51.6	45.1	46.1	45.9	46.9	56.8
United States	70.6	54.4	61.3	60.8	61.8	70.0
OECD	69.9	68.3	75.9	74.6	73.1	78.4

Source: www.oecd.org.

Ultimately, if the ratio of debt to GDP rises beyond a certain unknown point, then the solvency of the country comes into question. An additional indicator for potential insolvency is the ratio of interest rate payments to GDP, which is shown for some major economies in Exhibit 7-20. These represent payments required of governments to service their debts as a percentage of national output and as such reflect both the size of debts and the interest charged on them. Such ratios could rise rapidly with the growing debt ratios of 2009 and 2010, particularly if the interest rates on the debt were to rise from the historically low levels.

EXHIBIT 7-20 General Government Net Debt Interest Payments as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	3.5	1.7	1.0	0.8	0.6	0.5
Germany	2.9	2.7	2.4	2.4	2.4	2.3
Japan	1.3	1.5	0.8	0.6	0.6	0.9
United Kingdom	3.1	2.4	1.8	1.8	1.8	1.9
United States	3.5	2.5	1.8	1.8	1.9	1.8
OECD	3.6	2.5	1.8	1.7	1.7	1.6

Source: Organization for Economic Cooperation and Development (OECD).

Governments' spending was far in excess of revenues following the credit crisis of 2007–2010 as governments tried to stimulate their economies; this level of spending raised concerns in some quarters about the scale of governmental debt accumulation. Exhibit 7-19 shows that gross government financial liabilities relative to GDP for the OECD countries overall rose from 73.1 percent in 2007 to 78.4 percent in 2008. In Japan, where fiscal spending has been used to stimulate the economy from the early 1990s, the ratio has risen from 86.2 percent in 1995 to 172.1 percent in 2008. If an economy grows in real terms, so do the real tax revenues and hence the ability to service a growing real debt at constant tax rate levels. However, if the real growth in the economy is lower than the real interest rate on the debt, then the debt ratio will worsen even though the economy is growing, because the debt burden (i.e., the real interest rate times the debt) grows faster than the economy. Hence, an important issue for governments and their creditors is whether their additional spending leads to sufficiently higher tax revenues to pay the interest on the debt used to finance the extra spending.

However, within a national economy, the real value of the outstanding debt will fall if the overall price level rises (i.e., inflation, and hence a rise in nominal GDP even if real GDP is static) and thus the ratio of debt to GDP may not be rising. But if the general price level falls (i.e., deflation), then the ratio may stay elevated for a longer time. If net interest payments rise rapidly and investors lose confidence in a government's ability to honor its debts, then financing costs may escalate even more quickly and make the situation unstable.

Should we be concerned about the size of a national debt (relative to GDP)? There are strong arguments both for and against.

The arguments against being concerned about national debt (relative to GDP) are:

- The scale of the problem may be overstated because the debt is owed internally to fellow citizens. This is certainly the case in Japan, where 93 percent is owned by Japanese residents. In the United States and United Kingdom, the figures are 63 percent and 69 percent, respectively. South Korea and Canada have only 7 percent and 5 percent nonresident ownership of government debt, respectively. But Italy has 49 percent.[9]

[9]These data come from the Bank for International Settlements (BIS), International Monetary Fund (IMF), and central bank websites. All figures are as of 2008.

- A proportion of the money borrowed may have been used for capital investment projects or enhancing human capital (e.g., training, education); these should lead to raised future output and tax revenues.
- Large fiscal deficits require tax changes, which may actually reduce distortions caused by existing tax structures.
- Deficits may have no net impact, because the private sector may act to offset fiscal deficits by increasing saving in anticipation of future increased taxes. This argument is known as Ricardian equivalence and is discussed in more detail later.
- If there is unemployment in an economy, then the debt is not diverting activity away from productive uses (and indeed the debt could be associated with an increase in employment).

The arguments in favor of being concerned are:

- High levels of debt to GDP may lead to higher tax rates in the search for higher tax revenues. This may lead to disincentives to economic activity as the higher marginal tax rates reduce labor effort and entrepreneurial activity, leading to lower growth in the long run.
- If markets lose confidence in a government, then the central bank may have to print money to finance a government deficit. This may lead ultimately to high inflation, as evidenced by the economic history of Germany in the 1920s and more recently in Zimbabwe.
- Government borrowing may divert private-sector investment from taking place (an effect known as **crowding out**); if there is a limited amount of savings to be spent on investment, then larger government demands will lead to higher interest rates and lower private-sector investing.

An important distinction to make is between long- and short-run effects. Over short periods of time (say, a few years), crowding out may have little effect. If it lasts for a longer time, however, then capital accumulation in an economy may be damaged. Similarly, tax distortions may not be too serious over the short term but will have a more substantial impact over many years.

EXAMPLE 7-10 Types of Fiscal Policies

1. Which of the following is *not* associated with an expansionary fiscal policy?
   A. A rise in capital gains taxes
   B. Cuts in personal income taxes
   C. New capital spending by the government on road building
2. Fiscal expansions will *most likely* have the most impact on aggregate output when the economy is in which of the following states?
   A. Full employment
   B. Near full employment
   C. Considerable unemployment

3. Which one of the following is *most likely* a reason to *not* use fiscal deficits as an expansionary tool?
   A. They may crowd out private investment.
   B. They may facilitate tax changes to reduce distortions in an economy.
   C. They may stimulate employment when there is substantial unemployment in an economy.

Solution to 1: A is correct. A rise in capital gains taxes reduces income available for spending and hence reduces aggregate demand, other things being equal. Cutting income tax raises disposable income, while new road building raises employment and incomes; in both cases, aggregate demand rises and hence policy is expansionary.

Solution to 2: C is correct. When an economy is close to full employment, a fiscal expansion raising aggregate demand can have little impact on output because there are few spare unused resources (e.g., labor or idle factories); instead, there will be upward pressure on prices (i.e., inflation).

Solution to 3: A is correct. A frequent argument against raises in fiscal deficits is that the additional borrowing to fund the deficit in financial markets will displace private-sector borrowing for investment (i.e., crowd it out).

3.2. Fiscal Policy Tools and the Macroeconomy

We now look at the nature of the fiscal tools available to a government. Government spending can take a variety of forms:

- **Transfer payments** are welfare payments made through the social security system and, depending on the country, comprise payments for state pensions, housing benefits, tax credits and income support for poorer families, child benefits, unemployment benefits, and job search allowances. Transfer payments exist to provide a basic minimum level of income for low-income households, and they also provide a means by which a government can change the overall income distribution in a society. Note that these payments are not included in the definition of GDP because they do not reflect a reward to a factor of production for economic activity. Also, they are not considered to be part of general government spending on goods and services.
- **Current government spending** involves spending on goods and services that are provided on a regular, recurring basis—including health, education, and defense. Clearly, such spending will have a big impact on a country's skill level and overall labor productivity.
- **Capital expenditure** includes infrastructure spending on roads, hospitals, prisons, and schools. This investment spending will add to a nation's capital stock and affect productive potential for an economy.

Government spending can be justified on both economic and social grounds:

- To provide such services as defense that benefit all citizens equally.
- For infrastructure capital spending (e.g., roads) to help a country's economic growth.

- To guarantee a minimum level of income for poorer people and hence redistribute income and wealth (e.g., welfare and related benefits).
- To influence a government's economic objectives of low inflation and high employment and growth (e.g., management of aggregate demand).
- To subsidize the development of innovative and high-risk new products or markets (e.g., alternative energy sources).

Government revenues can take several forms:

- **Direct taxes** are levied on income, wealth, and corporate profits and include capital gains taxes, national insurance (or labor) taxes, and corporate taxes. They may also include a local income or property tax for both individuals and businesses. Inheritance tax on a deceased's estate will have both revenue-raising and wealth-redistribution aspects.
- **Indirect taxes** are taxes on spending on a variety of goods and services in an economy—such as the excise duties on fuel, alcohol, and tobacco, as well as sales tax (or value-added tax)—and often exclude health and education products on social grounds. In addition, taxes on gambling may also be considered to have a social aspect in deterring such activity, while fuel duties will have an environmental purpose in making fuel consumption and hence travel more expensive.

Taxes can be justified both in terms of raising revenues to finance expenditures and in terms of income and wealth redistribution policies. Economists typically consider four desirable attributes of a tax policy:

1. *Simplicity.* This refers to ease of compliance by the taxpayer and enforcement by the revenue authorities. The final liability should be certain and not easily manipulated.
2. *Efficiency.* Taxation should interfere as little as possible in the choices individuals make in the marketplace. Taxes affect behavior and should, in general, discourage work and investment as little as possible. A major philosophical issue among economists is whether tax policy should deliberately deviate from efficiency to promote good economic activities, such as savings, and discourage harmful ones, such as tobacco consumption. Although most would accept a limited role in guiding consumer choices, some will question how well equipped policy makers are to decide on such objectives and whether there will be unwanted ancillary effects, such as giving tax breaks for saving to people who already save and whose behavior does not change.
3. *Fairness.* This refers to the fact that people in similar situations should pay the same taxes (horizontal equity) and that richer people should pay more taxes (vertical equity). Of course, the concept of fairness is really subjective. Still, most would agree that income tax rates should be progressive—that is, that households and corporations should pay proportionately more as their incomes rise. However, some people advocate flat tax rates, whereby all should pay the same proportion of taxable income.
4. *Revenue sufficiency.* Although revenue sufficiency may seem obvious as a criterion for tax policy, there may be a conflict with fairness and efficiency. For example, one may believe that increasing income tax rates to reduce fiscal deficits reduces labor effort and that tax rate increases are thus an inefficient policy tool.

Some Issues with Tax Policy

1. *Incentives*. Some economists believe that income taxes reduce the incentive to work, save, and invest and that the overall tax burden has become excessive. These ideas are often associated with supply-side economics and the U.S. economist Arthur Laffer. A variety of income tax cuts and simplifications have taken place in the United States since 1981, and although there is substantial controversy, some claim that work effort did rise (although tax cuts had little impact on savings). Similarly, some found that business investment did rise, while others claimed it was independent of such cuts.
2. *Fairness*. How do we judge the fairness of the tax system? One way is to calibrate the tax burden falling on different groups of people ranked by their income and to assess how changes in taxes affect these groups. Of course, this imposes huge data demands on investigators and must be considered incomplete. In the United States, it has been found that the federal system is indeed highly progressive. Many countries use such methods to analyze the impact of tax changes on different income groups when they announce their annual fiscal policy plans.
3. *Tax reform*. There is continuous debate on reforming tax policy. Should there be a flat-rate tax on labor income? Should all investment be immediately deducted for corporate taxes? Should more revenue be sourced from consumption taxes? Should taxes be indexed to inflation? Should dividends be taxed when profits have already been subject to tax? Should estates be taxed at all? Many of these issues are raised in the context of their impact on economic growth.

EXAMPLE 7-11 Fiscal Tools

1. Which of the following is *not* a tool of fiscal policy?
   A. A rise in social transfer payments
   B. The purchase of new equipment for the armed forces
   C. An increase in deposit requirements for the buying of houses
2. Which of the following is *not* an indirect tax?
   A. Excise duty
   B. Value-added tax
   C. Employment taxes
3. Which of the following statements is *most* accurate?
   A. Direct taxes are useful for discouraging alcohol consumption.
   B. Because indirect taxes cannot be changed quickly, they are of no use in fiscal policy.
   C. Government capital spending decisions are slow to plan, implement, and execute and hence are of little use for short-term economic stabilization.

Solution to 1: C is correct. Rises in deposit requirements for house purchases are intended to reduce the demand for credit for house purchases and hence would be considered a tool of monetary policy. This is a policy used actively in several countries, and is under consideration by regulators in other countries to constrain house price inflation.

Solution to 2: C is correct. Both excise duty and VAT are applied to prices, whereas taxes on employment apply to labor income and hence are not indirect taxes.

Solution to 3: C is correct. Capital spending is much slower to implement than changes in indirect taxes, and indirect taxes affect alcohol consumption more directly than direct taxes.

3.2.1. The Advantages and Disadvantages of Using the Different Tools of Fiscal Policy

The different tools used to expedite fiscal policy as a means to try to put or keep an economy on a path of positive, stable growth with low inflation have both advantages and disadvantages:

Advantages

- Indirect taxes can be adjusted almost immediately after they are announced and can influence spending behavior instantly; they also generate revenue for the government at little or no cost to the government.
- Social policies, such as discouraging alcohol or tobacco use, can be adjusted almost instantly by raising such taxes.

Disadvantages

- Direct taxes are more difficult to change without considerable notice, often many months, because payroll computer systems will have to be adjusted (although the announcement itself may well have a powerful effect on spending behavior more immediately). The same may be said for welfare and other social transfers.
- Capital spending plans take longer to formulate and implement, typically over a period of years. For example, building a road or hospital requires detailed planning, legal permissions, and implementation. This is often a valid criticism of an active fiscal policy and was widely heard during the U.S. fiscal stimulus efforts in 2009–2010. On the other hand, such policies add to the productive potential of an economy, unlike a change in personal or indirect taxes. Of course, the slower the impact of a fiscal change, the more likely other exogenous changes will already be influencing the economy before the fiscal change kicks in.

The aforementioned tools may also have expectational effects at least as powerful as the direct effects. The announcement of future income tax rises a year ahead could potentially lead to reduced consumption immediately. Such delayed tax rises were a feature of U.K. fiscal policy of 2009–2010; however, the evidence is anecdotal because spending behavior changed little until the delayed tax changes actually came into force.

We may also consider the relative potency of the different fiscal tools. Direct government spending has a far bigger impact on aggregate spending and output than income tax cuts or transfer increases; however, if the latter are directed at the poorest in society (basically, those who spend all their income), then this will give a relatively strong boost. Further discussion and examples of these comparisons are given in Section 4 on the interaction between monetary and fiscal policy.

3.2.2. Modeling the Impact of Taxes and Government Spending: The Fiscal Multiplier

The conventional macroeconomic model has government spending, G, adding directly to aggregate demand, AD, and reducing it via taxes, T; these comprise both indirect taxes on expenditures and direct taxes on factor incomes. Further government spending is increased via the payment of transfer benefits, B, such as social security payments. Hence, the net impact of the government sector on aggregate demand is:

$$G - T + B = \text{Budget surplus OR deficit} \tag{7-5}$$

Net taxes (NT; taxes less transfers) reduce disposable income (YD) available to individuals relative to national income or output (Y) as follows:

$$YD = Y - NT = (1 - t)Y \tag{7-6}$$

where t is the **net tax rate**. Net taxes are often assumed to be proportional to national income, Y, and hence total tax revenue from net taxes is tY. If $t = 20\%$ or 0.2, then for every \$1 rise in national income, net tax revenue will rise by 20 cents and household disposable income will rise by 80 cents.

The **fiscal multiplier** is important in macroeconomics because it tells us how much output changes as exogenous changes occur in government spending or taxation. The recipients of the increase in government spending will typically save a proportion $1 - c$ of each additional dollar of disposable income, where c is the **marginal propensity to consume** (MPC) this additional income. Ignoring income taxes, we can see that $\$c$ will, in turn, be spent by these recipients on more goods and services. The recipients of this $\$c$ will themselves spend a proportion c of this additional income (i.e., $\$c \times c$, or c-squared). This process continues with income and spending growing at a constant rate of c as it passes from hand to hand through the economy. This is the familiar geometric progression, $1/(1 - c)$, with constant factor c, where $0 < c < 1$.

We define s as the **marginal propensity to save** (MPS), the amount saved out of an additional dollar of disposable income. Because $c + s = 1$, hence $s = 1 - c$.

In Exhibit 7-21, the MPC out of disposable income is 90 percent or 0.9 (72/80). The MPS is therefore $1 - 0.9$ or 0.1.

EXHIBIT 7-21 Disposable Income, Saving, and the Marginal Propensity to Consume

Income	Income Tax	Disposable Income	Consumption	Saving
\$100	\$20	\$80	\$72	\$8

For every dollar of new (additional) spending, total of incomes and spending rises by $\$1/(1 - c)$. And because $0 < c < 1$, this must be >1; this is the multiplier. If $c = 0.9$ (or individuals spend 90 percent of additions to income), then the multiplier $= 1/(1 - 0.9) = 10$.

A formal definition of the multiplier would be the ratio of the change in equilibrium output to the change in autonomous spending that caused the change. This is a monetary measure, but because prices are assumed to be constant in this analysis, real and monetary amounts are identical. Given that fiscal policy is about changes in government spending (G), net taxes (NT), and tax rates (t), we can see that the multiplier is an important tool for calibrating the possible impact of policy changes on output. How can we introduce tax changes into the multiplier concept? We do this by introducing the idea of disposable income, YD, defined as income less income taxes net of transfers, $Y - NT$.

Households spend a proportion c of disposable income (YD); that is, cYD or $c(Y - NT)$ or $c(1 - t)Y$. The marginal propensity to consume in the presence of taxes is then $c(1 - t)$. If the government increases spending, say on road building, by an amount G, then disposable income rises by $(1 - t)G$ and consumer spending by $c(1 - t)G$. Provided there are unused sources of capital and labor in the economy, this leads to a rise in aggregate demand and output; the recipients of this extra consumption spending will have $(1 - t)c(1 - t)G$ extra disposable income available and will spend c of it. This cumulative extra spending and income will continue to spread through the economy at a decreasing rate as $0 < c(1 - t) < 1$. The overall final impact on aggregate demand and output will effectively be the sum of this decreasing geometric series with common ratio $c(1 - t)$, and this sums to $1/[1 - c(1 - t)]$. This is known as the fiscal multiplier and is very relevant to studies of fiscal policy, as changes in G or tax rates will affect output in an economy through the value of the multiplier.

For example, if the tax rate is 20 percent, or 0.2, and the marginal propensity to spend is 90 percent, or 0.9, then the fiscal multiplier will be: $1/[1 - 0.9(1 - 0.2)]$ or $1/0.28 = 3.57$. In other words, if the government raises G by \$1 billion, total incomes and spending rise by \$3.57 billion.

Discretionary fiscal policy (see Section 3.3) will involve changes in these variables with a view to influencing Y.

3.2.3. The Balanced Budget Multiplier

If a government increases G by the same amount as it raises taxes, the aggregate output actually rises. Why is this?

It is because the marginal propensity to spend out of disposable income is less than 1, and hence for every dollar less in YD, spending falls only $\$c$. Hence, aggregate spending falls less than the tax rise by a factor of c. A balanced budget leads to a rise in output, which in turn leads to further rises in output and incomes via the multiplier effect.

Suppose an economy has an equilibrium output or income level of \$1,000 consisting of \$900 of consumption and \$100 of investment spending, which is fixed and not related to income. If government spending is set at \$200, financed by a tax rate of 20 percent (giving tax revenue of \$200), what will happen to output? First, additional government spending of \$200 will raise output by that amount; but will taxes of \$200 reduce output by a similar amount? Not if the MPC is less than 1; suppose it is 0.9, and hence spending will fall by only 90 percent of \$200, or \$180. The initial impact of the balanced fiscal package on aggregate demand will be to raise it by \$200 − \$180 = \$20. This additional output will, in turn, lead to further increases in income and output through the multiplier effect.

Even though this policy involved a combination of government spending and tax increases that initially left the government's budget deficit/surplus unchanged, the induced rise in output will lead to further tax revenue increases and a further change in the budget position. Could the government adjust the initial change in spending to offset exactly the eventual total change in tax revenues? The answer is yes, and we can ask: what will be the effect on output of this genuinely balanced budget change? This balanced budget multiplier always takes the value unity.

Government Debt, Deficits, and Ricardo

The total stock of government debt is the outstanding stock of IOUs issued by a government and not yet repaid. They are issued when the government has insufficient tax revenues to meet expenditures and has to borrow from the public. The size of the outstanding debt equals the cumulative quantity of net borrowing it has done, and the fiscal or budget deficit is added in the current period to the outstanding stock of debt. If the outstanding stock of debt falls, we have a negative deficit or a surplus.

If a government reduces taxation by $10 billion one year and replaces that revenue with borrowing of $10 billion from the public, will it have any real impact on the economy? The important issue here is how people perceive that action: Do they recognize what will happen over time as interest and bond principal have to be repaid out of future taxes? If so, they may think of the bond finance as equivalent to delayed taxation finance; thus, the reduction in current taxation will have no impact on spending because individuals save more in anticipation of higher future taxes to repay the bond. This is called **Ricardian equivalence** after the economist David Ricardo. If people do not correctly anticipate all the future taxes required to repay the additional government debt, then they feel wealthier when the debt is issued and may increase their spending, adding to aggregate demand.

Whether Ricardian equivalence holds in practice is ultimately an empirical issue and is difficult to calibrate conclusively given the number of things that are changing at any time in a modern economy.

3.3. Fiscal Policy Implementation: Active and Discretionary Fiscal Policy

In the following, we discuss major issues in fiscal policy implementation.

3.3.1. Deficits and the Fiscal Stance

An important question is the extent to which the budget is a useful measure of the government's fiscal stance. Does the size of the deficit actually indicate whether fiscal policy is expansionary or contractionary? Clearly, such a question is important for economic policy makers insofar as the deficit can change for reasons unrelated to actual fiscal policy changes.

EXHIBIT 7-22 General Government Net Cyclically Adjusted Borrowing or Lending as Percentage of GDP

	1995	2000	2005	2006	2007	2008
Australia	–3.1	0.1	1.1	1.4	1.3	0.1
Germany	–9.5	–1.8	–2.3	–1.5	–0.4	–0.5
Japan	–4.6	–7.1	–6.5	–1.8	–3.0	–2.3
United Kingdom	–5.6	0.9	–3.7	–3.3	–3.5	–5.1
United States	–2.9	0.7	–3.6	–2.6	–3.2	–6.1
OECD	–4.6	–1.1	–3.1	–1.9	–2.1	–3.7

Source: Organization for Economic Cooperation and Development (OECD).

For example, the automatic stabilizers mentioned earlier will lead to changes in the budget deficit unrelated to fiscal policy changes; a recession will cause tax revenues to fall and the budget deficit to rise. An observer may conclude that fiscal policy has been loosened and is expansionary and that no further government action is required.

To this end, economists often look at the **structural (or cyclically adjusted) budget deficit** as an indicator of the fiscal stance. This is defined as the deficit that would exist *if the economy was at full employment (or full potential output).* Hence, if we consider a period of relatively high unemployment, such as 2009–2010 with around 9 to 10 percent of the workforce out of work in the United States and Europe, then the budget deficits in those countries would be expected to be reduced substantially if the economies returned to full employment. At this level, tax revenues would be higher and social transfers lower. Recent data for major countries are given in Exhibit 7-22, where negative numbers refer to deficits and positive numbers are surpluses.

A further reason why actual government deficits may *not* be a good measure of fiscal stance is the distinction between real and nominal interest rates and the role of inflation adjustment when applied to budget deficits. Although national economic statistics treat the cash interest payments on debt as government expenditure, it makes more sense to consider only the inflation-adjusted (or real) interest payments because the real value of the outstanding debt is being eroded by inflation. Automatic stabilizers—such as income tax, VAT, and social benefits—are important because as output and employment fall and reduce tax revenues, *net* tax revenues also fall as unemployment benefits rise. This acts as a fiscal stimulus and serves to reduce the size of the multiplier, dampening the output response of whatever caused the fall in output in the first place. By their very nature, automatic stabilizers do not require policy changes; no policy maker has to decide that an economic shock has occurred and how to respond. Hence, the responsiveness of the economy to shocks is automatically reduced, as are movements in employment and output.

In addition to these automatic adjustments, governments also use discretionary fiscal adjustments to influence aggregate demand. These will involve tax changes and/or spending cuts or increases, usually with the aim of stabilizing the economy. A natural question is why fiscal policy cannot stabilize aggregate demand completely, hence ensuring full employment at all times.

3.3.2. Difficulties in Executing Fiscal Policy

Fiscal policy cannot stabilize aggregate demand completely, because the difficulties in executing fiscal policy cannot be completely overcome.

First, the policy makers do not have complete information on how the economy functions. It may take several months for policy makers to realize that an economy is slowing, because data appear with a considerable time lag and even then are subject to substantial revision. This is often called the **recognition lag** and has been likened to the problem of driving with the rearview mirror. Then, when policy changes are finally decided on, they may take many months to implement. This is the **action lag**. If a government decides to raise spending on capital projects to increase employment and incomes, for example, these may take many months to plan and put into action. Finally, the result of these actions on the economy will take additional time to become evident; this is the **impact lag**. These types of policy lags also occur in the case of discretionary monetary policy.

A second aspect of time in this process is the uncertainty of where the economy is heading independently of these policy changes. For example, a stimulus may occur simultaneously with a surprise rise in investment spending or in the demand for a country's exports just as discretionary government spending starts to rise. Macroeconomic forecasting models do not generally have a good track record for accuracy and hence cannot be relied on to aid the policy-making process in this context. In addition, when discretionary fiscal adjustments are announced (or are already underway), private-sector behavior may well change, leading to rises in consumption or investment, both of which will reinforce the effects of a rise in government expenditure. Again, this will make it difficult to calibrate the required fiscal adjustment to secure full employment.

There are wider macroeconomic issues also involved here:

- If the government is concerned with both unemployment *and* inflation in an economy, then raising aggregate demand toward the full employment level may also lead to a tightening labor market and rising wages and prices. The policy makers may be reluctant to further fine-tune fiscal policy in an uncertain world, because it might induce inflation.
- If the budget deficit is already large relative to GDP and further fiscal stimulus is required, then the necessary increase in the deficit may be considered unacceptable by the financial markets when government funding is raised, leading to higher interest rates on government debt and political pressure to tackle the deficit.
- Of course, all this presupposes that we know the level of full employment, which is difficult to measure accurately. Fiscal expansion raises demand, but what if we are already at full employment, which will be changing as productive capacity changes and workers' willingness to work at various wage levels changes?
- If unused resources reflect a low supply of labor or other factors rather than a shortage of demand, then discretionary fiscal policy will not add to demand and will be ineffective, raising the risk of inflationary pressures in the economy.
- The issue of crowding out may occur: If the government borrows from a limited pool of savings, the competition for funds with the private sector may crowd out private firms with subsequently less investing and economic growth. In addition, the cost of borrowing may rise, leading to the cancellation of potentially profitable opportunities. This concept is the subject of continuing empirical debate and investigation.

EXAMPLE 7-12 Evaluating Fiscal Policy

1. Which of the following statements is *least* accurate?
   A. The economic data available to policy makers have a considerable time lag.
   B. Economic models always offer an unambiguous guide to the future path of the economy.
   C. Surprise changes in exogenous economic variables make it difficult to use fiscal policy as a stabilization tool.
2. Which of the following statements is *least* accurate?
   A. Discretionary fiscal changes are aimed at stabilizing an economy.
   B. In the context of implementing fiscal policy, the recognition lag is often referred to as "driving with the rearview mirror."
   C. Automatic fiscal stabilizers include new plans for additional road building by the government.
3. Which of the following statements regarding a fiscal stimulus is *most* accurate?
   A. Accommodative monetary policy reduces the impact of a fiscal stimulus.
   B. Different statistical models will predict different impacts for a fiscal stimulus.
   C. It is always possible to predict precisely the impact of a fiscal stimulus on employment.
4. Which of the following statements is *most* accurate?
   A. An increase in the budget deficit is always expansionary.
   B. An increase in government spending is always expansionary.
   C. The structural deficit is always larger than the deficit below full employment.
5. Crowding out refers to:
   A. a fall in interest rates that reduces private investment.
   B. a rise in private investment that reduces private consumption.
   C. a rise in government borrowing that reduces the ability of the private sector to access investment funds.
6. A contractionary fiscal policy will always involve which of the following?
   A. A balanced budget
   B. A reduction in government spending
   C. A fall in the budget deficit or rise in the surplus
7. Which one of the following statements is *most* accurate?
   A. Ricardian equivalence refers to individuals having no idea of future tax liabilities.
   B. If there is high unemployment in an economy, then easy monetary and fiscal policies should lead to an expansion in aggregate demand.
   C. Governments do not allow political pressures to influence fiscal policies but do allow voters to affect monetary policies.

Solution to 1: B is correct. Economic forecasts from models will always have an element of uncertainty attached to them and thus are not unambiguous or precise in their prescriptions. Once a fiscal policy decision has been made and implemented, unforeseen changes in other variables may affect the economy in ways that would lead to

changes in the fiscal policy if we had perfect foresight. Note that it is true that official economic data may be available with substantial time lags, making fiscal judgments more difficult.

Solution to 2: C is correct. New plans for road building are discretionary and not automatic.

Solution to 3: B is correct. Different models embrace differing views on how the economy works, including differing views on the impact of fiscal stimuli.

Solution to 4: A is correct. Note that increases in government spending may be accompanied by even bigger rises in tax receipts and hence may not be expansionary.

Solution to 5: C is correct. A fall in interest rates is likely to lead to a rise in investment. Crowding out refers to government borrowing that reduces the ability of the private sector to invest.

Solution to 6: C is correct. Note that a reduction in government spending could be accompanied by an even bigger fall in taxation, making it expansionary.

Solution to 7: B is correct. Note that governments often allow pressure groups to affect fiscal policy and that Ricardian equivalence involves individuals correctly anticipating future taxes, so A and C are not correct choices.

4. THE RELATIONSHIP BETWEEN MONETARY AND FISCAL POLICY

Both monetary and fiscal policies can be used to try to influence the macroeconomy. But the impact of monetary policy on aggregate demand may differ depending on the fiscal policy stance. Conversely, the impact of fiscal policy might vary under various alternative monetary policy conditions. Clearly, policy makers need to understand this interaction. For example, they need to consider the impact of changes to the budget when monetary policy is accommodative as opposed to when it is restrictive: Can we expect the same impact on aggregate demand in both situations?

Although both fiscal and monetary policy can alter aggregate demand, they do so through differing channels with differing impact on the composition of aggregate demand. The two policies are not interchangeable. Consider the following cases in which the assumption is made that *wages and prices are rigid*:

- *Easy fiscal policy/tight monetary policy.* If taxes are cut or government spending rises, the expansionary fiscal policy will lead to a rise in aggregate output. If this is accompanied by a reduction in money supply to offset the fiscal expansion, then interest rates will rise and have a negative effect on private-sector demand. We have higher output and higher interest rates, and government spending will be a larger proportion of overall national income.

- *Tight fiscal policy/easy monetary policy.* If a fiscal contraction is accompanied by expansionary monetary policy and low interest rates, then the private sector will be stimulated and will rise as a share of GDP, while the public sector will shrink.
- *Easy monetary policy/easy fiscal policy.* If both fiscal and monetary policy are easy, then the joint impact will be highly expansionary—leading to a rise in aggregate demand, lower interest rates (at least if the monetary impact is larger), and growing private and public sectors.
- *Tight monetary policy/tight fiscal policy.* Interest rates rise (at least if the monetary impact on interest rates is larger) and reduce private demand. At the same time, higher taxes and falling government spending lead to a drop in aggregate demand from both public and private sectors.

4.1. Factors Influencing the Mix of Fiscal and Monetary Policy

Although governments are concerned about stabilizing the level of aggregate demand at close to the full employment level, they are also concerned with the growth of potential output. To this end, encouraging private investment will be important. It may best be achieved by accommodative monetary policy with low interest rates and a tight fiscal policy to ensure free resources for a growing private sector.

At other times, the lack of a good quality, trained workforce or perhaps a modern capital infrastructure will be seen as an impediment to growth; thus, an expansion in government spending in these areas may be seen as a high priority. If taxes are not raised to pay for this, then the fiscal stance will be expansionary. If a loose monetary policy is chosen to accompany this expansionary spending, then it is *possible* that inflation may be induced. Of course, it is an open question as to whether policy makers can judge the appropriate levels of interest rates or fiscal spending levels.

Clearly, the mix of policies will be heavily influenced by the political context. A weak government may raise spending to accommodate the demands of competing vested interests (e.g., subsidies to particular sectors, such as agriculture), and thus a restrictive monetary policy may be needed to hold back the possibly inflationary growth in aggregate demand through raised interest rates and less credit availability.

Both fiscal and monetary policies suffer from lack of precise knowledge of where the economy is today, because data appear initially subject to revision and with a time lag. However, fiscal policy suffers from two further issues with regard to its use in the short run.

As we saw earlier, it is difficult to implement quickly because spending on capital projects takes time to plan, procure, and put into practice. In addition, it is politically easier to loosen fiscal policy than to tighten it; in many cases, automatic stabilizers are the source of fiscal tightening, because tax rates are not changing and political opposition is muted. Similarly, the independence of many central banks means that decisions on raising interest rates are outside the hands of politicians and thus can be made more easily.

The interaction between monetary and fiscal policies was also implicitly evident in our discussion of Ricardian equivalence because if tax cuts have no impact on private spending as individuals anticipate future higher taxes, then clearly this may lead policy makers to favor monetary tools.

Ultimately, the interaction of monetary and fiscal policies in practice is an empirical question, which we touched on earlier. In their detailed research paper using the IMF's Global Integrated Monetary and Fiscal Model (International Monetary Fund 2009), IMF researchers

examined four forms of coordinated global fiscal loosening over a two-year period, which will be reversed gradually after the two years are completed. These are:

1. An increase in social transfers to all households.
2. A decrease in tax on labor income.
3. A rise in government investment expenditure.
4. A rise in transfers to the poorest in society.

The two types of monetary policy responses considered are:

1. No monetary accommodation, so rising aggregate demand leads to higher interest rates immediately.
2. Interest rates are kept unchanged (accommodative policy) for the two years.

The following important policy conclusions from this study emphasize the role of policy interactions:

- *No monetary accommodation.* Government spending increases have a much bigger effect (six times bigger) on GDP than similar size social transfers because the latter are not considered permanent, although real interest rates rise as monetary authorities react to rises in aggregate demand and inflation. Targeted social transfers to the poorest citizens have double the effect of the nontargeted transfers, while labor tax reductions have a slightly bigger impact than the nontargeted social transfers.
- *Monetary accommodation.* Except for the case of the cut in labor taxes, fiscal multipliers are now much larger than when there is no monetary accommodation. The cumulative multiplier (i.e., the cumulative effect on real GDP over the two years divided by the percentage of GDP, which is a fiscal stimulus) is now 3.9 for government expenditure compared to 1.6 with no monetary accommodation. The corresponding numbers for targeted social transfer payments are 0.5 without monetary accommodation and 1.7 with it. The larger multiplier effects with monetary accommodation result from rises in aggregate demand and inflation, leading to falls in real interest rates and additional private-sector spending (e.g., on investment goods). Labor tax cuts are less positive.

4.2. Quantitative Easing and Policy Interaction

What about the scenario of zero interest rates and deflation? Fiscal stimulus should still raise demand and inflation, lowering real interest rates and stimulating private-sector demand. We saw earlier that quantitative easing has been a feature of major economies during 2009 and 2010. This involves the purchase of government or private securities by the central bank from individuals, institutions, or banks and substituting central bank balances for those securities. The ultimate aim is that recipients will subsequently increase expenditures, lending or borrowing in the face of raised cash balances and lower interest rates.

If the central bank purchases government securities on a large scale, it is effectively funding the budget deficit, and the independence of monetary policy is an illusion. This so-called printing of money is feared by many economists as the monetization of the government deficit. Note that it is unrelated to the conventional inflation target of central banks, such as the Bank of England. Some economists question whether an independent central bank should engage in such activity.

4.3. The Importance of Credibility and Commitment

The IMF model implies that if governments run persistently high budget deficits, real interest rates rise and crowd out private investment, reducing each country's productive potential. As individuals realize that deficits will persist, inflation expectations and longer-term interest rates rise. This reduces the effect of the stimulus by half.

Further, if there were a real lack of commitment to fiscal discipline over the longer term (e.g., because of aging populations) and the ratio of government debt to GDP rose by 10 percentage points permanently in the United States alone, then world real interest rates would rise by 0.14 percent—leading to a 0.6 percent permanent fall in world GDP.

EXAMPLE 7-13 Interactions of Monetary and Fiscal Policy

1. In a world where Ricardian equivalence holds, governments would *most likely* prefer to use monetary rather than fiscal policy because under Ricardian equivalence:
   A. real interest rates have a more powerful effect on the real economy.
   B. the transmission mechanism of monetary policy is better understood.
   C. the future impact of fiscal policy changes is fully discounted by economic agents.
2. If fiscal policy is easy and monetary policy tight, then:
   A. interest rates would tend to fall, reinforcing the fiscal policy stance.
   B. the government sector would tend to shrink as a proportion of total GDP.
   C. the government sector would tend to expand as a proportion of total GDP.
3. Which of the following has the greatest impact on aggregate demand according to an IMF study? A 1 percent of GDP stimulus in:
   A. government spending.
   B. rise in transfer benefits.
   C. cut in labor income tax across all income levels.

Solution to 1: C is correct. If Ricardian equivalence holds, then economic agents anticipate that the consequence of any current tax cut will be future tax rises, which leads them to increase their saving in anticipation of this so that the tax cut has little effect on consumption and investment decisions. Governments would be forced to use monetary policy to affect the real economy on the assumption that money neutrality did not hold in the short term.

Solution to 2: C is correct. With a tight monetary policy, real interest rates should rise and reduce private-sector activity, which could be at least partially offset by an expansion in government activity via the loosening of fiscal policy. The net effect, however, would be an expansion in the size of the public sector relative to the private sector.

Solution to 3: A is correct. The study clearly showed that direct spending by the government leads to a larger impact on GDP than changes in taxes or benefits.

5. SUMMARY

In this chapter, we have sought to explain the practices of both monetary and fiscal policy. Both can have a significant impact on economic activity, and it is for this reason that financial analysts need to be aware of the tools of both monetary and fiscal policy, the goals of the monetary and fiscal authorities, and, most important, the monetary and fiscal policy transmission mechanisms.

- Governments can influence the performance of their economies by using combinations of monetary and fiscal policy. Monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy. By contrast, fiscal policy refers to the government's decisions about taxation and spending. The two sets of policies affect the economy via different mechanisms.
- Money fulfills three important functions: It acts as a medium of exchange, provides individuals with a way of storing wealth, and provides society with a convenient unit of account. Via the process of fractional reserve banking, the banking system can create money.
- The amount of wealth that the citizens of an economy choose to hold in the form of money—as opposed to, for example, bonds or equities—is known as the demand for money. There are three basic motives for holding money: transactions-related, precautionary, and speculative.
- The addition of one unit of additional reserves to a fractional reserve banking system can support an expansion of the money supply by an amount equal to the money multiplier, defined as 1/reserve requirement (stated as a decimal).
- The nominal rate of interest is comprised of three components: a real required rate of return, a component to compensate lenders for future inflation, and a risk premium to compensate lenders for uncertainty (e.g., about the future rate of inflation).
- Central banks take on multiple roles in modern economies. They are usually the monopoly supplier of their currency, the lender of last resort to the banking sector, the government's bank, and the banks' bank, and they often supervise banks. Although they may express their objectives in different ways, the overarching objective of most central banks is price stability.
- For a central bank to be able to implement monetary policy objectively, it should have a degree of independence from government, be credible, and be transparent in its goals and objectives.
- The ultimate challenge for central banks as they try to manipulate the supply of money to influence the economy is that they cannot control the amount of money that households and corporations put in banks on deposit, nor can they easily control the willingness of banks to create money by expanding credit. Taken together, this also means that they cannot always control the money supply. Therefore, there are definite limits to the power of monetary policy.
- The concept of money neutrality is usually interpreted as meaning that money cannot influence the real economy in the long run. However, by the setting of its policy rate, a central bank hopes to influence the real economy via the policy rate's impact on other market interest rates, asset prices, the exchange rate, and the expectations of economic agents.
- Inflation targeting is the most common monetary policy—although exchange rate targeting is also used, particularly in developing economies. Quantitative easing attempts to spur aggregate demand by drastically increasing the money supply.
- Fiscal policy involves the use of government spending and revenue raising (taxation) to impact a number of aspects of the economy: the overall level of aggregate demand in an economy and hence the level of economic activity, the distribution of income and wealth among different segments of the population, and hence ultimately the allocation of resources among different sectors and economic agents.

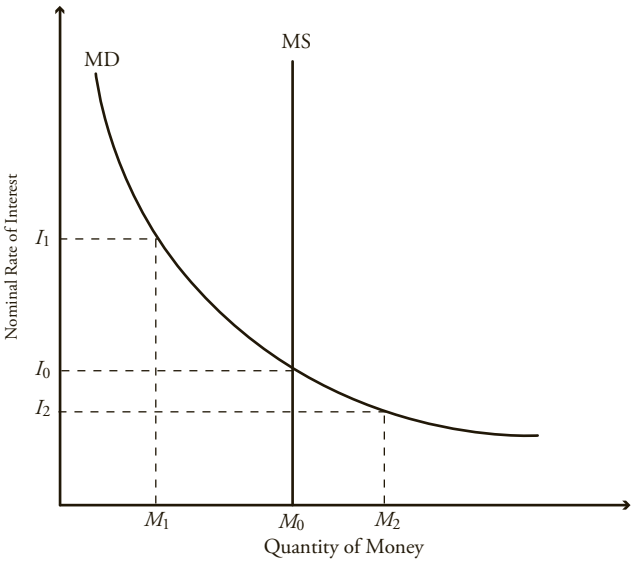
- The tools that governments use in implementing fiscal policy are related to the way in which they raise revenue and the different forms of expenditure. Governments usually raise money via a combination of direct and indirect taxes. Government expenditure can be current on goods and services or can take the form of capital expenditure—for example, on infrastructure projects.
- As economic growth weakens or when it is in recession, a government can enact an expansionary fiscal policy—for example, by raising expenditures without an offsetting increase in taxation. Conversely, by reducing expenditures and maintaining tax revenues, a contractionary policy might reduce economic activity. Fiscal policy can therefore play an important role in stabilizing an economy.
- Although both fiscal and monetary policy can alter aggregate demand, they work through different channels; the policies are therefore not interchangeable, and they conceivably can work against one another unless the government and central bank coordinate their objectives.

PRACTICE PROBLEMS[10]

1. As the reserve requirement increases, the money multiplier:
   A. increases.
   B. decreases.
   C. remains the same.

2. Which is the *most* accurate statement regarding the demand for money?
   A. Precautionary money demand is directly related to GDP.
   B. Transactions money demand is inversely related to returns on bonds.
   C. Speculative demand is inversely related to the perceived risk of other assets.

3. The following exhibit shows the supply and demand for money:



[10]These practice problems were developed by Karen O'Connor Rubsam, CFA (Fountain Hills, Arizona, USA).

There is an excess supply of money when the nominal rate of interest is:
A. I_0.
B. I_1.
C. I_2.

4. According to the theory of money neutrality, money supply growth does *not* affect variables such as real output and employment in:
   A. the long run.
   B. the short run.
   C. the long run and the short run.

5. Which of the following *best* describes a fundamental assumption when monetary policy is used to influence the economy?
   A. Financial markets are efficient.
   B. Money is not neutral in the short run.
   C. Official rates do not affect exchange rates.

6. Monetarists are *most likely* to believe that:
   A. there is a causal relationship running from inflation to money.
   B. inflation can be affected by changing the money supply growth rate.
   C. rapid financial innovation in the market increases the effectiveness of monetary policy.

7. The proposition that the real interest rate is relatively stable is *most* closely associated with:
   A. the Fisher effect.
   B. money neutrality.
   C. the quantity theory of money.

8. Which of the following equations is a consequence of the Fisher effect?
   A. Nominal interest rate = Real interest rate + Expected rate of inflation
   B. Real interest rate = Nominal interest rate + Expected rate of inflation
   C. Nominal interest rate = Real interest rate + Market risk premium

9. Central banks would typically be *most* concerned with costs of:
   A. low levels of inflation that are anticipated.
   B. moderate levels of inflation that are anticipated.
   C. moderate levels of inflation that are not anticipated.

10. Monetary policy is *least likely* to include:
    A. setting an inflation rate target.
    B. changing an official interest rate.
    C. enacting a transfer payment program.

11. Which role is a central bank *least likely* to assume?
    A. Lender of last resort
    B. Sole supervisor of banks
    C. Supplier of the currency

12. Which is the *most* accurate statement regarding central banks and monetary policy?
    A. Central bank activities are typically intended to maintain price stability.
    B. Monetary policies work through the economy via four independent channels.
    C. Commercial and interbank interest rates move inversely to official interest rates.

13. When a central bank announces a decrease in its official policy rate, the desired impact is an increase in:
    A. investment.
    B. interbank borrowing rates.
    C. the national currency's value in exchange for other currencies.

14. Which action is a central bank *least likely* to take if it wants to encourage businesses and households to borrow for investment and consumption purposes?
    A. Sell long-dated government securities.
    B. Purchase long-dated government Treasuries.
    C. Purchase mortgage bonds or other securities.

15. A central bank that decides the desired levels of interest rates and inflation and the horizon over which the inflation objective is to be achieved is *most* accurately described as being:
    A. target independent and operationally independent.
    B. target independent but not operationally independent.
    C. operationally independent but not target independent.

16. A country that maintains a target exchange rate is *most likely* to have which outcome when its inflation rate rises above the level of the inflation rate in the target country?
    A. An increase in short-term interest rates
    B. An increase in the domestic money supply
    C. An increase in its foreign currency reserves

17. A central bank's repeated open market purchases of government bonds:
    A. decrease the money supply.
    B. are prohibited in most countries.
    C. are consistent with an expansionary monetary policy.

18. In theory, setting the policy rate equal to the neutral interest rate should promote:
    A. stable inflation.
    B. a balanced budget.
    C. greater employment.

19. A prolonged period of an official interest rate of zero without an increase in economic growth *most likely* suggests that:
    A. quantitative easing must be limited to be successful.
    B. there may be limits to the effectiveness of monetary policy.
    C. targeting reserve levels is more important than targeting interest rates.

20. Raising the reserve requirement is *most likely* an example of which type of monetary policy?
    A. Neutral
    B. Expansionary
    C. Contractionary

21. Which of the following is a limitation on the ability of central banks to stimulate growth in periods of deflation?
    A. Ricardian equivalence
    B. The interaction of monetary and fiscal policy
    C. The fact that interest rates have a minimum value (zero percent)

22. The *least likely* limitation to the effectiveness of monetary policy is that central banks cannot:
    A. accurately determine the neutral rate of interest.
    B. regulate the willingness of financial institutions to lend.
    C. control amounts that economic agents deposit into banks.

23. Which of the following is the *most likely* example of a tool of fiscal policy?
    A. Public financing of a power plant
    B. Regulation of the payment system
    C. Central bank's purchase of government bonds

24. The *least likely* goal of a government's fiscal policy is to:
    A. redistribute income and wealth.
    B. influence aggregate national output.
    C. ensure the stability of the purchasing power of its currency.

25. Given an independent central bank, monetary policy actions are *more likely* than fiscal policy actions to be:
    A. implementable quickly.
    B. effective when a specific group is targeted.
    C. effective when combating a deflationary economy.

26. Which statement regarding fiscal policy is *most* accurate?
    A. To raise business capital spending, personal income taxes should be reduced.
    B. Cyclically adjusted budget deficits are appropriate indicators of fiscal policy.
    C. An increase in the budget surplus is associated with expansionary fiscal policy.

27. The *least likely* explanation for why fiscal policy cannot stabilize aggregate demand completely is that:
    A. private-sector behavior changes over time.
    B. policy changes are implemented very quickly.
    C. fiscal policy focuses more on inflation than on unemployment.

28. Which of the following *best* represents a contractionary fiscal policy?
    A. Public spending on a high-speed railway
    B. A temporary suspension of payroll taxes
    C. A freeze in discretionary government spending

29. A pay-as-you-go rule, which requires that any tax cut or increase in entitlement spending be offset by an increase in other taxes or reduction in other entitlement spending, is an example of which fiscal policy stance?
    A. Neutral
    B. Expansionary
    C. Contractionary

30. Quantitative easing, the purchase of government or private securities by the central banks from individuals and institutions, is an example of which monetary policy stance?
    A. Neutral
    B. Expansionary
    C. Contractionary

31. The *most likely* argument against high national debt levels is that:
    A. the debt is owed internally to fellow citizens.
    B. they create disincentives for economic activity.
    C. they may finance investment in physical and human capital.

32. Which statement regarding fiscal deficits is *most* accurate?
    A. Higher government spending may lead to higher interest rates and lower private-sector investing.
    B. Central bank actions that grow the money supply to address deflationary conditions decrease fiscal deficits.
    C. According to the Ricardian equivalence, deficits have a multiplicative effect on consumer spending.

33. Which policy alternative is *most likely* to be effective for growing both the public and private sectors?
    A. Easy fiscal policy/easy monetary policy
    B. Easy fiscal policy/tight monetary policy
    C. Tight fiscal policy/tight monetary policy

CHAPTER 8

INTERNATIONAL TRADE AND CAPITAL FLOWS

Usha Nair-Reichert

Daniel Robert Witschi, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Compare gross domestic product and gross national product.
- Describe the benefits and costs of international trade.
- Distinguish between comparative advantage and absolute advantage.
- Explain the Ricardian and Heckscher–Ohlin models of trade and the source(s) of comparative advantage in each model.
- Compare types of trade and capital restrictions and their economic implications.
- Explain motivations for and advantages of trading blocs, common markets, and economic unions.
- Describe the balance of payments accounts, including their components.
- Explain how decisions by consumers, firms, and governments affect the balance of payments.
- Describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund, and the World Trade Organization.

1. INTRODUCTION

Global investors must address two fundamentally interrelated questions: where to invest and in what asset classes? Some countries may be attractive from an equity perspective because of their strong economic growth and the profitability of particular domestic sectors or industries. Other countries may be attractive from a fixed-income perspective because of their interest rate environment and price stability. To identify markets that are expected to provide

attractive investment opportunities, investors must analyze cross-country differences in such factors as expected gross domestic product (GDP) growth rates, monetary and fiscal policies, trade policies, and competitiveness. From a longer-term perspective investors also need to consider such factors as a country's stage of economic and financial market development, demographics, quality and quantity of physical and human capital (accumulated education and training of workers), and its area(s) of comparative advantage.[1]

This chapter provides a framework for analyzing a country's trade and capital flows and their economic implications. International trade can facilitate economic growth by increasing the efficiency of resource allocation, providing access to larger capital and product markets, and facilitating specialization based on comparative advantage. The flow of financial capital (funds available for investment) between countries with excess savings and those where financial capital is scarce can increase liquidity, raise output, and lower the cost of capital. From an investment perspective, it is important to understand the complex and dynamic nature of international trade and capital flows because investment opportunities are increasingly exposed to the forces of global competition for markets, capital, and ideas.

This chapter is organized as follows. Section 2 defines basic terminology used in the chapter and describes patterns and trends in international trade and capital flows. It also discusses the benefits of international trade, distinguishes between absolute and comparative advantage, and explains two traditional models of comparative advantage. Section 3 describes trade restrictions and their implications and discusses the motivation for, and advantages of, trade agreements. Section 4 describes the balance of payments, and Section 5 discusses the function and objectives of international organizations that facilitate trade. A summary of key points and practice problems conclude the chapter.

2. INTERNATIONAL TRADE

The following sections describe the role, importance, and possible benefits and costs of international trade. Before beginning those discussions, we define some basic terminology used in this area.

2.1. Basic Terminology

The aggregate output of a nation over a specified time period is usually measured as its gross domestic product or its gross national product. Gross domestic product (GDP) measures the market value of all final goods and services produced by factors of production (such as labor and capital) located within a country/economy during a given period of time, generally a year or a quarter. Gross national product (GNP), however, measures the market value of all final goods and services produced by factors of production (such as labor and capital) supplied by residents of a country, regardless of whether such production takes place within the country or outside of the country. Two differences between a country's GDP and its GNP are that GDP includes, whereas GNP excludes, the production of goods and services by foreigners within that country, and GNP includes, whereas GDP excludes, the production of goods and services by its citizens outside of the country. Countries that have large differences between GDP and

[1]Comparative advantage refers to a country's ability to produce a good at a lower cost than other goods it produces, as compared with another country. It will be more precisely defined and illustrated in Section 2.4.

GNP generally have a large number of citizens who work abroad (for example, Pakistan and Portugal), and/or pay more for the use of foreign-owned capital in domestic production than they earn on the capital they own abroad (for example, Brazil and Canada). Therefore, GDP is more widely used as a measure of economic activity occurring *within* the country, which, in turn, affects employment, growth, and the investment environment.

Imports are goods and services that a domestic economy (i.e., households, firms, and government) purchases from other countries. For example, the U.S. economy imports (purchases) cloth from India and wine from France. **Exports** are goods and services that a domestic economy sells to other countries. For example, South Africa exports (sells) diamonds to the Netherlands, and China exports clothing to the European Union. So how are services imported or exported? If a Greek shipping company transports the wine that the United States imports from France, the United States would classify the cost of shipping as an import of services from Greece and the wine would be classified as an import of goods from France. Similarly, when a British company provides insurance coverage to a South African diamond exporter, Britain would classify the cost of the insurance as an export of services to South Africa. Other examples of services exported or imported include engineering, consulting, and medical services.

The **terms of trade** are defined as the ratio of the price of exports to the price of imports, representing those prices by export and import price indexes, respectively. The terms of trade capture the relative cost of imports in terms of exports. If the prices of exports increase relative to the prices of imports, the terms of trade have improved because the country will be able to purchase more imports with the same amount of exports.[2] For example, when oil prices increased during 2007–2008, major oil exporting countries experienced an improvement in their terms of trade because they had to export less oil in order to purchase the same amount of imported goods. In contrast, if the price of exports decreases relative to the price of imports, the terms of trade have deteriorated because the country will be able to purchase fewer imports with the same amount of exports. Because each country exports and imports a large number of goods and services, the terms of trade of a country are usually measured as an index number (normalized to 100 in some base year) that represents a ratio of the average price of exported goods and services to the average price of imported goods and services. Exhibit 8-1 shows the terms of trade reported in Salvatore (2010). A value over 100 indicates that the country, or group of countries, experienced better terms of trade relative to the base year of 2000; a value under 100 indicates worse terms of trade than in 2000.

As an example, Exhibit 8-1 indicates that from 1990 to 2006 both of the broader groups, developing and industrial countries, experienced a slight decline in their terms of trade. Looking at the disaggregated data indicates that developing countries in Asia and the western hemisphere experienced a considerable decline in terms of trade whereas those in Europe and the Middle East (which benefited from rising prices of petroleum exports) experienced a substantial increase. Africa also experienced a small improvement in its terms of trade during this period.

Net exports is the difference between the value of a country's exports and the value of its imports (i.e., value of exports minus imports). If the value of exports equals the value of imports, then trade is balanced. If the value of exports is greater than the value of imports,

[2]Although the prices of imports and exports are each stated in currency units, the currency units cancel out when we take the ratio, so the terms of trade reflect the relative price of imports and exports in real (i.e., quantity) terms: units of imports per unit of exports. To see this, note that if one unit of imports costs P_M currency units and one unit of exports is priced at P_X currency units, then the country can buy P_X/P_M (= Terms of trade) units of imports for each unit of exports.

EXHIBIT 8-1 Data on the Terms of Trade for Industrial and Developing Countries (Unit Export Value/Unit Import Value)

	1990	1995	2000	2005	2006
Industrial countries	99.8	104.8	100	101.3	99.0
Developing countries	103.0	101.9	100	99.4	100.5
Africa	100.4	102.8	100	107.9	105.2
Asia	106.8	106.8	100	91.5	89.2
Europe	68.7	105.5	100	102.1	99.8
Middle East	109.0	68.4	100	140.4	155.9
Western hemisphere	129.6	107.1	100	104.3	108.7

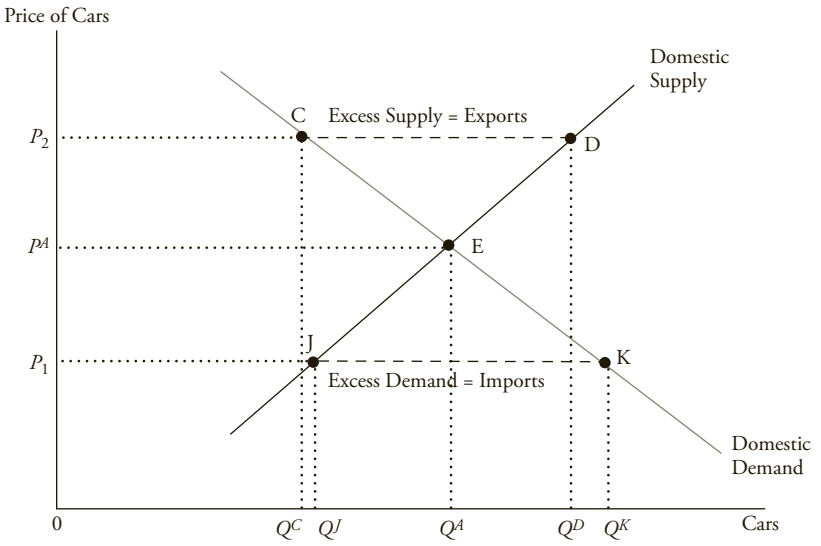
Source: Salvatore (2010), case study 3-3. Base year 2000 = 100.

then there is a **trade surplus**; if the value of exports is less, there is a **trade deficit**. When a country has a trade surplus, it lends to foreigners or buys assets from foreigners reflecting the financing needed by foreigners running trade deficits with that country. Similarly, when a country has a trade deficit, it has to borrow from foreigners or sell some of its assets to foreigners. Section 4 on the balance of payments explains these relationships more fully.

Autarky is a state in which a country does not trade with other countries. This means that all goods and services are produced and consumed domestically. The price of a good or service in such an economy is called its **autarkic price**. An autarkic economy is also known as a **closed economy** because it does not trade with other countries. An **open economy**, in contrast, is an economy that does trade with other countries. If there are no restrictions on trade, then members of an open economy can buy and sell goods and services at the price prevailing in the world market, the world price. An open economy can provide domestic households with a larger variety of goods and services, give domestic companies access to global markets and customers, and offer goods and services that are more competitively priced. In addition, it can offer domestic investors access to foreign capital markets, foreign assets, and greater investment opportunities. For capital-intensive industries, such as automobiles and aircraft, manufacturers can take advantage of economies of scale because they have access to a much larger market. **Free trade** occurs when there are no government restrictions on a country's ability to trade. Under free trade, global aggregate demand and supply determine the equilibrium quantity and price of imports and exports. Government policies that impose restrictions on trade, such as tariffs and quotas (discussed later in the chapter), are known as **trade protection** and prevent market forces (demand and supply) from determining the equilibrium price and quantity for imports and exports. According to Alan Deardorff, *globalization* refers to the "increasing worldwide integration of markets for goods, services, and capital that began to attract special attention in the late 1990s."[3] It also references "a variety of other changes that were perceived to occur at about the same time, such as an increased role for large corporations (multinational corporations) in the world economy and increased

[3]Alan Deardorff, "Deardorff's Glossary of International Economics," www-personal.umich.edu/~alandear/glossary.

EXHIBIT 8-2 Excess Demand, Excess Supply, Imports, and Exports



intervention into domestic policies and affairs by international institutions," such as the International Monetary Fund, the World Trade Organization, and the World Bank.

The levels of aggregate demand and supply and the quantities of imports and exports in an economy are related to the concepts of *excess demand* and *excess supply*. Exhibit 8-2 shows supply and demand curves for cars in the United Kingdom. E is the autarkic equilibrium at price P^A and quantity Q^A, with the quantity of cars demanded equaling the quantity supplied. Now, consider a situation in which the country opens up to trade and the world price is P_1. At this price, the quantity demanded domestically is Q^K while the quantity supplied is Q^J. Hence excess demand is Q^JQ^K. This quantity is satisfied by imports. For example, at a world price of $15,000, the quantity of cars demanded in the United Kingdom might be 2.0 million and UK production of cars only 1.5 million. As a result, the excess demand of 500,000 would be satisfied by imports. Returning to Exhibit 8-2, now consider a situation in which the world price is P_2. The quantity demanded is Q^C while the quantity supplied is Q^D. Hence, the domestic excess supply at world price P_2 is Q^CQ^D, which results in exports of Q^CQ^D.

2.2. Patterns and Trends in International Trade and Capital Flows

The importance of trade in absolute and relative terms (trade to GDP ratio) is illustrated in Exhibits 8-3 through 8-5. Exhibit 8-3 shows that trade as a percentage of regional GDP increased in all regions of the world during 1970–2006. Developing countries in Asia had the fastest growth in trade, increasing from less than 20 percent of GDP in 1970 to more than 90 percent of GDP in 2006.

Exhibit 8-4 indicates that trade as a percentage of GDP and the GDP growth rate have increased in most regions of the world during 1990–2006. However, data for 2008 (not shown) indicate a decline that, although consistent with the worldwide economic downturn, varied across country groups. High-income countries that are members of the Organization

EXHIBIT 8-3 Trade in Goods and Services (Percentage of Regional GDP)

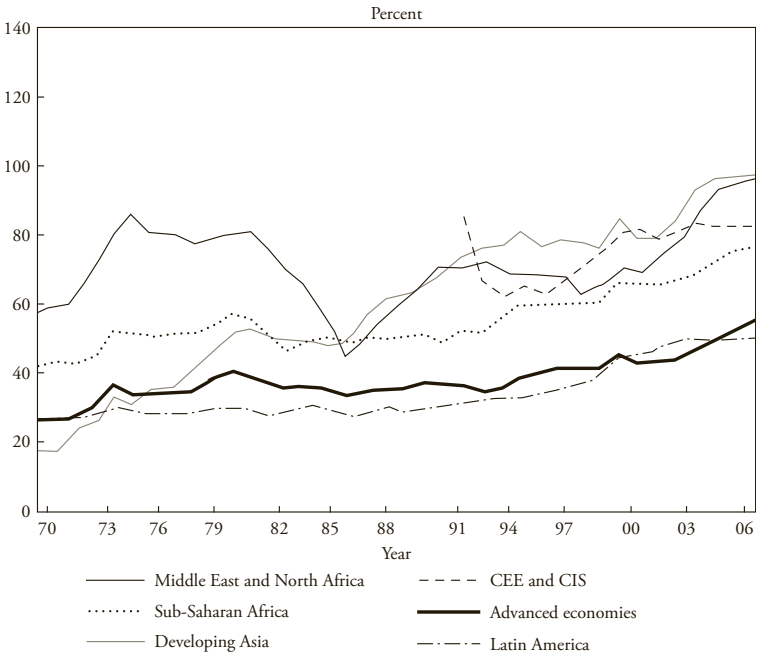


EXHIBIT 8-4 Trade Openness and GDP Growth

	Trade as Percentage of GDP (averaged over the period)			Average GDP Growth (%)		
Country Group	1980–1989	1990–1999	2000–2006	1980–1989	1990–1999	2000–2006
World	37.2	41.0	50.7	3.1	2.7	3.2
High income:						
All	38.1	40.3	49.5	3.1	2.6	2.5
OECD	35.3	37.2	44.7	3.1	2.5	2.4
Non-OECD	120.0	128.1	172.5	3.9	4.5	5.0
Low and middle income:						
All	32.4	44.4	56.9	3.4	3.5	5.8
Middle	32.4	44.5	57.1	3.4	3.5	5.8
Upper middle	33.4	44.3	53.5	2.1	1.7	4.1
Lower middle	31.4	44.8	61.4	6.0	6.1	7.7
Low	32.5	39.9	51.7	2.6	2.7	4.8

Note: Averages indicate the average of the annual data for the period covered.
Source: World Bank.

for Economic Cooperation and Development (OECD) experienced a growth rate of 2.4 percent during 2000–2006, but had a growth rate of only 0.3 percent in 2008. The corresponding numbers for growth in non-OECD high-income countries are 5.0 percent and 3.2 percent, respectively; for lower-middle-income countries, they are 7.7 percent and 7.5 percent, respectively. The 2009 *World Development Report* affirmed the link between trade and growth and noted evidence that all rich and emerging economies are oriented to being open to trade. More specifically, the report indicated:

> When exports are concentrated in labor-intensive manufacturing, trade increases the wages for unskilled workers, benefiting poor people. It also encourages macroeconomic stability, again benefiting the poor, who are more likely to be hurt by inflation. And through innovation and factor accumulation, it enhances productivity and thus growth. There may be some empirical uncertainty about the strength of trade's relationship with growth. But essentially all rich and emerging economies have a strong trade orientation. (World Bank 2009)

Of course, trade is not the only factor that influences economic growth. Research has also identified such factors as the quality of institutions, infrastructure, and education; economic systems; the degree of development; and global market conditions (World Trade Organization 2008).

Exhibit 8-5 presents trade and foreign direct investment as a percentage of GDP for select countries for 1980–2007. **Foreign direct investment** (FDI) refers to direct investment by a firm in one country (the *source country*) in productive assets in a foreign country (the *host*

EXHIBIT 8-5 Increasing Global Interdependence: Foreign Direct Investment and Trade as a Percentage of GDP

Country/Region	Type of Flow	1980	1990	2000	2007
World	Trade	38.4	38.0	48.8	57.3
	FDI: Net inflows	0.6	1.0	5.1	4.3
	FDI: Net outflows	0.6	1.1	3.6	4.5
Argentina	Trade	11.5	15.0	22.4	45.0
	FDI: Net inflows	−0.1	0.0	0.3	0.6
	FDI: Net outflows	0.9	1.3	3.7	2.5
Germany	Trade	45.3	49.7	66.4	86.7
	FDI: Net inflows	0.5	1.4	3.1	4.9
	FDI: Net outflows	0.0	0.2	11.1	2.3
India	Trade	15.6	15.7	27.4	45.2
	FDI: Net inflows	0.0	0.0	0.1	1.4
	FDI: Net outflows	0.0	0.1	0.8	2.0
United States	Trade	20.8	20.5	25.9	28.7
	FDI: Net inflows	0.7	0.6	1.6	2.9
	FDI: Net outflows	0.6	0.8	3.2	1.9

Source: World Development Indicators, World Bank.

country). When a firm engages in FDI, it becomes a **multinational corporation** (MNC) operating in more than one country or having subsidiary firms in more than one country. It is important to distinguish FDI from **foreign portfolio investment** (FPI), which refers to shorter-term investment by individuals, firms, and institutional investors (e.g., pension funds) in such foreign financial instruments as foreign stocks and foreign government bonds. Exhibit 8-5 shows that trade as a percentage of GDP for the world as a whole increased from 38 percent in 1980 to 57 percent in 2007. In Argentina, trade as a percentage of GDP increased from 11.5 percent in 1980 to 45.0 percent in 2007, while in India during this same period it increased from 15.6 percent to 45.2 percent. Among the more advanced economies, trade expanded sharply in Germany (from 45.3 percent to 86.7 percent), but in the United States trade expanded more modestly (from 20.8 percent to 28.7 percent).

The increasing importance of multinational corporations is also apparent in Exhibit 8-5. Net FDI inflows and outflows increased as a percentage of GDP between 1980 and 2000 for each of the countries shown. Trade between multinational firms and their subsidiaries (i.e., intrafirm trade) has become an important part of world trade. For example, 46 percent of U.S. imports occur between related parties (Bernard, Jensen, Redding, and Schott 2010). Globalization of production has increased the productive efficiency of manufacturing firms because they are able to decompose their value chains into individual components or parts, and then outsource their production to different countries where these components can be produced most efficiently.[4] For example, Nintendo's Wii remote is manufactured with components sourced from several countries in the world: the accelerometer is manufactured in the United States; the base memory chip in Italy; and the data converter in the United States, Thailand, and India; the plastic casing is assembled in China and designed in Japan; the Bluetooth chip is manufactured in Taiwan and designed in the United States (California); and the rumble pack is manufactured in various countries in Asia.[5] Foreign direct investment and outsourcing have increased business investment in these countries and provided smaller and less developed countries the opportunity to participate in international trade. For example, the *World Investment Report* (United Nations 2002) indicates that in January 2002 Intel had 13 fabrication plants and 11 assembly and testing sites in seven countries. It was the leading national exporter from Ireland, the Philippines, and Costa Rica, and 17th among foreign exporters from China. These trends indicate the increasing global interdependence of economies, although the degree of interdependence varies among regions and countries. Greater interdependence also means that countries are now more exposed to global competition. As a result they must be more flexible in their production structure in order to respond effectively to changes in global demand and supply.

The complexity of trading relationships has also increased with the development of sophisticated global supply chains that include not only final goods but also intermediate goods and services. Increased global interdependence has changed the risk and return profiles of many countries. Countries that have greater international links are more exposed to, and

[4]Hill (2007, 412) explains the idea of the firm as a value chain: "The operations of the firms can be thought of as a value chain composed of a series of distinct value creation activities including production, marketing and sales, materials management, R&D, human resources, information systems, and firm infrastructure." Production itself can be broken down into distinct components and each component outsourced separately.

[5]http://money.cnn.com/magazines/fortune/storysupplement/wiiremote/index.htm.

affected by, economic downturns and crises occurring in other parts of the world. The contagion effect of the Asian financial crisis, which began in Thailand in July 1997, spread to many other markets, such as Indonesia, Malaysia, South Korea, the Philippines, Hong Kong, Singapore, and Taiwan. It even affected Brazil and Russia to some degree, although there is less clarity about the mechanisms by which the crisis spread beyond Asia. Among the outward symptoms of the crisis were exchange rate problems, such as currency speculation and large depreciation of currencies, capital flight, and financial and industrial sector bankruptcies. However, recovery was surprisingly swift, and all of these countries exhibited positive growth by the second quarter of 1999 (Gerber 2010).

2.3. Benefits and Costs of International Trade

The preceding sections have described the growth of world trade and the increasing interdependence of national economies. Has trade been beneficial? The benefits and costs of international trade have been widely debated. The most compelling arguments supporting international trade are: countries gain from exchange and specialization, industries experience greater economies of scale, households and firms have greater product variety, competition is increased, and resources are allocated more efficiently.

Gains from exchange occur when trade enables each country to receive a higher price for its exports (and greater profit) and/or pay a lower price for imported goods instead of producing these goods domestically at a higher cost (i.e., less efficiently). This exchange, in turn, leads to a more efficient allocation of resources by increasing production of the export good and reducing production of the import good in each country (trading partner). This efficiency allows consumption of a larger bundle of goods, thus increasing overall welfare. The fact that trade increases overall welfare does not, of course, mean that every individual consumer and producer is better off. What it does mean is that the winners could, in theory, compensate the losers and still be better off.

Trade also leads to greater efficiency by fostering specialization based on comparative advantage. Traditional trade models, such as the Ricardian model and the Heckscher–Ohlin model, focus on specialization and trade according to comparative advantage arising from differences in technology and factor endowments, respectively. These models are discussed in the next section.

Newer models of trade focus on the gains from trade that result from economies of scale, greater product variety, and increased competition. In an open economy, increased competition from foreign firms reduces the monopoly power of domestic firms and forces them to become more efficient, as compared to a closed economy. Industries that exhibit increasing returns to scale (for example, the automobile and steel industries) benefit from increased market size as a country starts trading, because the average cost of production declines as output increases in these industries. Monopolistically competitive models of trade have been used to explain why there is significant two-way trade (known as *intra-industry trade*) between countries within the same industry. Intra-industry trade occurs when a country exports and imports goods in the same product category or classification.

In a monopolistically competitive industry, there are many firms; each firm produces a unique or differentiated product, there are no exit or entry barriers, and long-run economic profits are zero. In such a model, even though countries may be similar, they gain from trade because each country focuses on the production and export of one or more varieties of the good and imports other varieties of the good. For example, the European Union exports and imports different types of cars. Consumers gain from having access to a greater variety of final

goods. Firms benefit from greater economies of scale because firms both within and outside the EU are able to sell their goods in both markets. Hence, scale economies allow firms to benefit from the larger market size and to experience lower average cost of production as a result of trade.

Research suggests that trade liberalization can lead to increased real (that is, inflation-adjusted) GDP although the strength of this relationship is still debated. The positive influence of trade on GDP can arise from more efficient allocation of resources, learning by doing, higher productivity, knowledge spillovers, and trade-induced changes in policies and institutions that affect the incentives for innovation.[6] In industries where there is learning by doing, such as the semiconductor industry, the cost of production per unit declines as output increases because of expertise and experience acquired in the process of production. Trade can lead to increased exchange of ideas, freer flow of technical expertise, and greater awareness of changing consumer tastes and preferences in global markets. It can also contribute to the development of higher-quality and more effective institutions and policies that encourage domestic innovation. For example, Coe and Helpman (1995) show that foreign research and development (R&D) has beneficial effects on domestic productivity. These effects become stronger the more open a country is to foreign trade. They estimate that about a quarter of the benefits of R&D investment in a G-7 country accrues to its trading partners.[7] Hill (2007) discusses the case of Logitech, a Swiss company that manufactures computer mice. In order to win original equipment manufacturer (OEM) contracts from IBM and Apple, Logitech needed to develop innovative designs and provide high-volume production at a low cost. So in the late 1980s the company moved to Taiwan, which had a highly qualified labor force, competent parts suppliers, and a rapidly expanding local computer industry, and offered Logitech space in a science park at a very competitive rate. Soon thereafter, Logitech was able to secure the Apple contract.

Opponents of free trade point to the potential for greater income inequality and the loss of jobs in developed countries as a result of import competition. As a country moves toward free trade, there will be adjustments in domestic industries that are exporters as well as those that face import competition. Resources (investments) may need to be reallocated into or out of an industry depending on whether that industry is expanding (exporters) or contracting (i.e., facing import competition). As a result of this adjustment process, less efficient firms may be forced to exit the industry, which may, in turn, lead to higher unemployment and the need for displaced workers to be retrained for jobs in expanding industries. The counterargument is that although there may be short-term and even some medium-term costs, these resources are likely to be more effectively (re)employed in other industries in the long run. Nonetheless, the adjustment process is virtually certain to impose costs on some groups of stakeholders. For example, the U.S. textile industry has undergone significant changes over the past 30 years as a result of competition from lower-priced imports produced in developing countries, including increased outsourcing of production by U.S. firms. Example 8-1 discusses recent developments and projections for future employment in the industry.

[6]Knowledge spillovers occur when investments in knowledge creation generate benefits that extend beyond the investing entity and facilitate learning and innovation by other firms or entities.

[7]G-7 countries include Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States.

EXAMPLE 8-1 The U.S. Textile Industry

According to the U.S. Bureau of Labor Statistics (BLS), the textile, textile product, and apparel manufacturing industry is very labor intensive and faces strong import competition. Changing trade regulations have a big impact on employment in this industry. In 2005, members of the World Trade Organization terminated the Multifibre Arrangement that imposed quotas for apparel and textile products. This agreement included most U.S. trading partners and, in particular, China. The expiration of quotas in 2005 has allowed more apparel and textile products to be imported into the United States. Although some bilateral quotas have been reimposed between the United States and China, imports have increased substantially. The low-skilled, labor-intensive parts of the industry, such as fabric for apparel, have lost many jobs as firms shift their operations to countries with very cheap labor costs. The more skill-intensive jobs, such as design jobs and custom or high-end items that are produced domestically, have not been as adversely affected by trade. Firms in the highly automated and innovative sectors of the industry, such as industrial fabrics, carpets, and specialty yarns, are competitive on a global scale, and it is expected that they will increase their exports as a result of the liberalization of trade in textiles.

The BLS estimates presented in Exhibit 8-6 indicate that in 2008 there were 497,100 wage and salary workers in the textile, textile product, and apparel manufacturing industries. California, Georgia, and North Carolina together accounted for about 44 percent of these workers. The BLS also estimates that there will be a 47.9 percent decline in employment in this sector between 2008 and 2018. Increased labor productivity and foreign competition will both continue to contribute to this trend.

The segment of the industry that remains in the United States has responded to competitive pressures by adopting new and more advanced technologies and becoming very labor efficient. Advanced technology includes computers and computer-controlled equipment that aid in many functions, such as design, pattern making, and pattern

EXHIBIT 8-6 Employment in Textile, Textile Product, and Apparel Manufacturing by Industry Segment

2008 and Projected Change 2008–2018 (employment in thousands)		
Industry Segment	2008 Employment	2008–2018 Percentage Change
Textile, textile product, and apparel manufacturing	497.1	−47.9
Textile mills	151.1	−47.6
Textile product mills	147.6	−38.1
Apparel manufacturing	198.4	−55.4

Source: BLS National Employment Matrix, 2008–2018.

cutting; wider looms; and the use of robotics to move material within the plant. All these initiatives are boosting productivity, providing workers with increased training and new skills, and changing the nature of work in the industry. Among the domestic industry's advantages are its proximity to the domestic market and its ability to respond to fashion trends more rapidly than foreign competitors can. The domestic industry is also better positioned to participate in retailers' move to just-in-time inventory management systems and electronic data exchange systems.

1. What are the key changes in trade policy that have affected the U.S. textile, textile product, and apparel manufacturing industry?
2. How did increased import competition affect the U.S. industry?

Solution to 1: The Multifibre Arrangement that imposed quotas for apparel and textile products came to an end in 2005. This affected trade with most U.S. trading partners and, in particular, China. The expiration of quotas in 2005 has allowed more apparel and textile products to be imported into the United States. Although some bilateral quotas have been reimposed between the United States and China, imports have increased substantially.

Solution to 2: One of the main impacts was that many low-skilled workers lost their jobs as firms moved production out of the United States to lower-wage countries. The affect on more skill-intensive jobs, such as those involved with design and domestically produced custom and high-end items, has been less severe. The highly automated and innovative sectors of the industry, such as industrial fabrics, carpets, and specialty yarns, are competitive on a global scale. The industry has responded to competitive pressures by adopting new and advanced technologies and becoming very labor efficient.

EXAMPLE 8-2 Benefits of Trade

Consider two countries that each produce two goods. Suppose the cost of producing cotton relative to lumber is lower in Cottonland than in Lumberland.

1. How would trade between the two countries affect the lumber industry in Lumberland?
2. How would trade between the two countries affect the lumber industry in Cottonland?
3. What would happen to the lumber industry workers in Cottonland in the long run?
4. What is the meaning of the expression *gains from trade*?
5. What are some of the benefits from trade?

Solution to 1: The lumber industry in Lumberland would benefit from trade. Because the cost of producing lumber relative to producing cotton is lower in Lumberland than in Cottonland (i.e., lumber is relatively cheap in Lumberland), Lumberland will export lumber and the industry will expand.

Solution to 2: The lumber industry in Cottonland would not benefit from trade, at least in the short run. Because lumber is relatively expensive to produce in Cottonland, the domestic lumber industry will shrink as lumber is imported from Lumberland.

Solution to 3: The overall welfare effect in both countries is positive. However, in the short run, many lumber producers in Cottonland (and cotton producers in Lumberland) are likely to find themselves without jobs as the lumber industry in Cottonland and the cotton industry in Lumberland contract. Those with skills that are also needed in the other industry may find jobs fairly quickly. Others are likely to do so after some retraining. In the long run, displaced workers should be able to find jobs in the expanding export industry. However, those who remain in the import-competing industry may be permanently worse off because their industry-specific skills are now less valuable. Thus, even in the long run, trade does not necessarily make every stakeholder better off. But the winners could compensate the losers and still be better off, so the overall welfare effect of opening trade is positive.

Solution to 4: Gains from trade imply that the overall benefits of trade outweigh the losses from trade. It does not mean that all stakeholders (producers, consumers, government) benefit (or benefit equally) from trade.

Solution to 5: Some of the benefits from trade are gains from exchange and specialization based on relative cost advantage, gains from economies of scale as the companies add new markets for their products, greater variety of products available to households and firms, greater efficiency from increased competition, and more efficient allocation of resources.

2.4. Comparative Advantage and the Gains from Trade

Up to this point, we have not been precise about what it means for a country to have a comparative advantage in the production of specific goods and services. In this section, we define comparative advantage, distinguish it from the notion of absolute advantage, and demonstrate the gains from trading in accordance with comparative advantage. We then explain two traditional models of trade—the Ricardian and Heckscher–Ohlin models—and the source of comparative advantage in each model.

2.4.1. Gains from Trade: Absolute and Comparative Advantage

A country has an **absolute advantage** in producing a good (or service) if it is able to produce that good at a lower cost or use fewer resources in its production than its trading partner. For example, suppose a worker in Brazil can produce either 20 pens or 40 pencils in a day. A worker in China can produce either 10 pens or 60 pencils. A Chinese worker produces 60 pencils a day while a Brazilian worker produces only 40 pencils a day. Hence, China produces pencils at a lower cost than Brazil, and has an absolute advantage in the production of pencils. Similarly, Brazil produces pens at a lower cost than China, and hence has an absolute advantage in the production of pens. A country has a **comparative advantage** in producing a good if its opportunity cost of producing that good is less than that of its trading partner. In our example, the opportunity cost of producing an extra pen in China is six pencils. It is the

opportunity forgone—namely, the number of pencils China would have to give up to produce an extra pen. If Brazil does not trade and has to produce both pens and pencils, it will have to give up two pencils in order to produce a pen. Similarly, in China each pen will cost six pencils. Hence, the opportunity cost of a pen in Brazil is two pencils, whereas in China it is six pencils. Brazil has the lower opportunity cost and thus a comparative advantage in the production of pens. China has a lower opportunity cost (one pencil costs one-sixth of a pen) than Brazil (one pencil costs half a pen) in the production of pencils and thus has a comparative advantage in the production of pencils. Example 8-3 further illustrates these concepts.

EXAMPLE 8-3 Absolute and Comparative Advantages

Suppose there are only two countries, India and the United Kingdom. India exports cloth to the United Kingdom and imports machinery. The output per worker per day in each country is shown in Exhibit 8-7.

EXHIBIT 8-7 Output per Worker per Day

	Machinery	Cloth (yards)
United Kingdom	**4**	8
India	2	**16**

Based only on the information given, address the following:

1. Which country has an absolute advantage in the production of:
   A. machinery?
   B. cloth?
2. Do the countries identified in Question 1 as having an absolute advantage in the production of (A) machinery and (B) cloth also have a comparative advantage in those areas?

Solution to 1A: The United Kingdom has an absolute advantage in the production of machinery because it produces more machinery per worker per day than India.

Solution to 1B: India has an absolute advantage in the production of cloth because it produces more cloth per worker per day than the United Kingdom.

Solution to 2: In both cases, the answer is yes. In the case of machinery, the opportunity cost of a machine in the United Kingdom is 2 yards of cloth (8 ÷ 4 or 1 machine = 2 yards of cloth). This amount is the autarkic price of machines in terms of cloth in the United Kingdom. In India, the opportunity cost of a machine is 8 yards of cloth (16 ÷ 2 or 1 machine = 8 yards of cloth). Thus, the United Kingdom has a comparative advantage in producing machines. In contrast, the opportunity cost of a yard of cloth in the United Kingdom and in India is one-half and one-eighth of a machine, respectively. India has a lower opportunity cost (one-eighth of a machine) and, therefore, a comparative advantage in the production of cloth.

It is important to note that even if a country does not have an absolute advantage in producing any of the goods, it can still gain from trade by exporting the goods in which it has a comparative advantage. In Example 8-3, if India could produce only six yards of cloth per day instead of 16 yards of cloth, the United Kingdom would have an *absolute* advantage in both machines and cloth. However, India would still have a *comparative* advantage in the production of cloth because the opportunity cost of a yard of cloth in India, one-third of a machine in this case, would still be less than the opportunity cost of a yard of cloth in the United Kingdom (one-half of a machine as before).

Let us now illustrate the gains from trading according to comparative advantage. In Example 8-3, if the United Kingdom could sell a machine for more than two yards of cloth and if India could purchase a machine for less than eight yards of cloth, both countries would gain from trade. Although it is not possible to determine the exact world price without additional details regarding demand and supply conditions, both countries would gain from trade as long as the world price for machinery in terms of cloth is between the autarkic prices of the trading partners. In our example, this price corresponds to a price of between two and eight yards of cloth for a machine. *The further away the world price of a good or service is from its autarkic price in a given country, the more that country gains from trade.* For example, if the United Kingdom were able to sell a machine to India for seven yards of cloth (i.e., closer to India's autarkic price), it would gain five yards of cloth per machine sold to India compared with its own autarkic price (with no trade) of one machine for two yards of cloth. However, if the United Kingdom were able to sell a machine to India for only three yards of cloth (closer to the UK autarkic price), it would gain only one yard of cloth per machine sold to India compared with its own autarkic price.

Exhibits 8-8 and 8-9 provide the production and consumption schedules of both countries at autarky and after trade has commenced. In autarky (Exhibit 8-8), the United Kingdom produces and consumes 200 machines and 400 yards of cloth (without trade, consumption of each product must equal domestic production). Similarly, India produces 100 machines and 800 yards of cloth in autarky. In a world economy consisting of only these two

EXHIBIT 8-8 Production and Consumption in Autarky

	Autarkic Production	Autarkic Consumption
United Kingdom		
Machinery (m)	200	200
Cloth (yards) (c)	400	400
India		
Machinery	100	100
Cloth (yards)	800	800
Total World		
Machinery	300	300
Cloth (yards)	1,200	1,200

EXHIBIT 8-9 Gains from Trade

	Posttrade Production	Posttrade Consumption	Change in Consumption (compared with autarky)
United Kingdom			
Machinery	400	240	+40
Cloth (yards)	0	640	+240
India			
Machinery	0	160	+60
Cloth (yards)	1,600	960	+160
Total World			
Machinery	400	400	+100
Cloth (yards)	1,600	1,600	+400

countries, total output for each commodity is the sum of production in both countries. Therefore, total world output is 300 machines and 1,200 yards of cloth.

Now, assume that the United Kingdom and India start trading and that the world price of one machine is four yards of cloth (1m = 4c). This price is within the range of acceptable world trading prices discussed earlier because this price lies between the autarkic prices of the United Kingdom (1m = 2c) and India (1m = 8c). Exhibit 8-9 shows that in an open economy, the United Kingdom would specialize in machines and India would specialize in cloth. As a result, the United Kingdom produces 400 machines and no cloth, while India produces 1,600 yards of cloth and no machines. The United Kingdom exports 160 machines to India in exchange for 640 yards of cloth. After trade begins with India, the United Kingdom consumes 240 machines and 640 yards of cloth. Consumption in the United Kingdom increases by 40 machines and 240 yards of cloth. Similarly, India consumes 160 machines and 960 yards of cloth, an increase of 60 machines and 160 yards of cloth. World production and consumption is now 400 machines and 1,600 yards of cloth. Posttrade production and consumption exceeds the autarkic situation by 100 machines and 400 yards of cloth.

Exhibit 8-10a and Exhibit 8-10b show a more general case of gains from trade under increasing costs. In Exhibit 8-10a, the curve connecting the *x*-axis and *y*-axis is the UK production possibilities frontier (PPF).[8] That is, it represents the combinations of cloth and machinery that the United Kingdom can produce given its technology and resources (capital and labor). The slope of the PPF at any point is the opportunity cost of one good in terms of the other. The shape of the PPF indicates increasing opportunity cost in terms of machines as more cloth is produced and vice versa. To maximize the value of output, production occurs where the slope of the PPF equals the relative price of the goods. P^A represents the autarkic price line, which is tangent to the PPF at point A, the autarkic equilibrium. The slope of the

[8]Modified from Salvatore (2010).

EXHIBIT 8-10a Graphical Depiction of Gains from Trade with Increasing Costs: United Kingdom

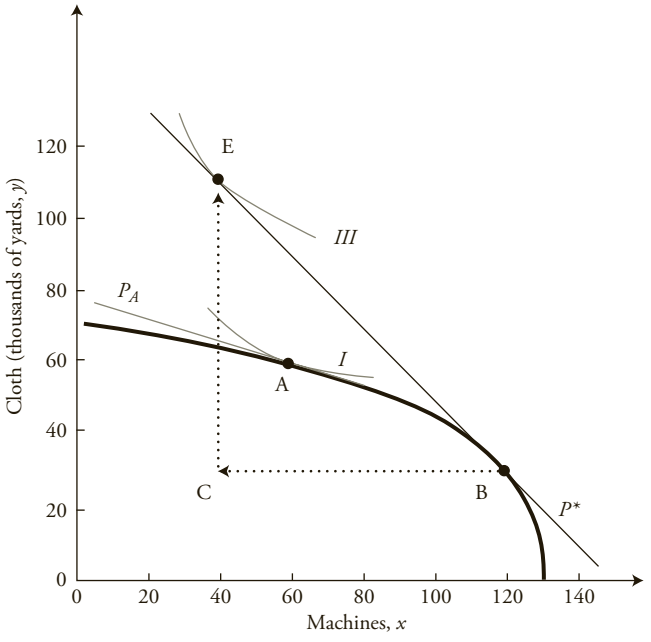
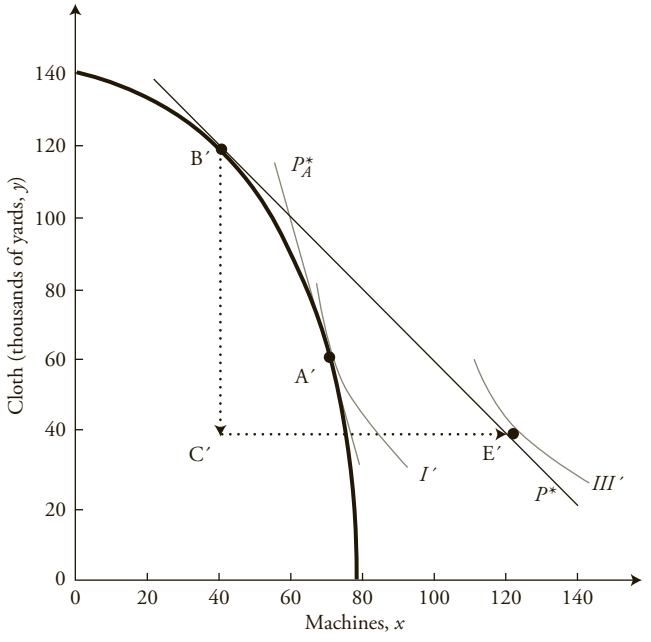


EXHIBIT 8-10b Graphical Depiction of Gains from Trade with Increasing Costs: India



autarkic price line represents the opportunity cost before trade. In autarky, the United Kingdom produces and consumes 60 machines and 60 thousand yards of cloth, and is on indifference curve *I*.[9] When the United Kingdom starts trading with India, it faces the world price line *P**. This new price line is tangent to the PPF at point B. The change in relative prices of the goods encourages the United Kingdom to increase the production of the good in which it has comparative advantage (machines) and produce at point B instead of point A. We note that at point B the United Kingdom has increased the production of machines to 120 units and reduced the production of cloth to 30 thousand yards. We also note that trade has expanded the UK consumption possibilities. The United Kingdom consumes at point E after trade, exports 80 machines to India, and imports 80 thousand yards of cloth from India. Note that point E is outside the PPF, but on the world price line that is tangent to the PPF at point B. This line is also the trading possibilities line because trade occurs along this line. The slope of this line is the opportunity cost of a machine in terms of cloth in the world market. The United Kingdom has clearly increased its welfare through trade because it is able to consume at point E, which is on a higher indifference curve (*III*) and thus represents a higher level of welfare compared with the autarkic consumption point A on indifference curve *I*.

Exhibit 8-10b shows the corresponding situation for India. When trade opens with the United Kingdom, India shifts production from point A′ to point B′, producing more cloth, the good in which it has a comparative advantage, and fewer machines. It now exports 80 thousand yards of cloth to the United Kingdom and imports 80 machines from the United Kingdom. India now consumes at E′, which is on the world price line and also on a higher indifference curve, *III*′, than the autarkic consumption point (A′) on indifference curve *I*′. Thus, by specializing (incompletely, as is typically the case with increasing production costs) in the good in which it has a comparative advantage, each country increases its welfare. We should also note that *P** is the price at which trade is balanced. At this relative world price, the export of cloth from India equals the import of cloth into the United Kingdom (80 thousand yards) and the export of machines from the United Kingdom equals the imports of machines into India (80 machines).

A country's comparative advantage can change over time as a result of structural shifts in its domestic economy, shifts in the global economy, the accumulation of physical or human capital, new technology, the discovery of such natural resources as oil, and so on. For example, an increase in skilled labor in China has led several multinational companies to establish R&D facilities in China to benefit from its highly educated workforce.

EXAMPLE 8-4 Changes in Comparative Advantage

Exhibit 8-11 shows how Taiwan's comparative advantage changed over time as a result of an export-oriented development strategy it adopted during the 1960s.* The challenges of foreign competition created a virtuous circle that was self-reinforcing. Taiwan's changing comparative advantage was the result of government policy, an increasingly skilled and productive workforce, and proactive firms that learned and adapted new technology.

[9]An indifference curve represents the various combinations of goods (machines and cloth) that provide the same level of utility or welfare. Higher indifference curves represent higher levels of utility or welfare.

EXHIBIT 8-11 Changes in Structure of Taiwan's Merchandise Exports, 1963–2003 (Percentage Shares)

Products	1963	1973	1983	1993	2003
Agricultural products	59.3	15.4	8.0	5.1	2.5
Mining products	2.7	0.9	2.4	1.7	3.5
Manufactures	38.0	83.6	89.1	93.0	93.7
Iron and steel	3.0	1.3	2.5	1.6	3.9
Chemicals	5.1	1.5	2.4	5.1	8.1
Other semimanufactures	11.7	12.1	11.6	9.6	6.8
Machinery and transport equipment	1.5	23.5	26.2	44.4	55.7
Office and telecom equipment	0.3	16.3	13.9	23.8	35.8
Electrical machinery and apparatus	0.3	2.7	3.6	6.5	8.1
Textiles	11.7	12.8	7.2	9.6	6.2
Clothing	3.0	16.1	11.9	4.4	1.4
Other consumer goods	1.8	16.3	27.4	18.4	11.6

Source: World Trade Organization (2008).

1. How has Taiwan's structure of exports changed over time?
2. How did increased foreign competition impact the economy?
3. What were the factors that helped to change Taiwan's comparative advantage?

Solution to 1: The economy moved from exporting agricultural products and textiles during the 1960s to exporting clothes and other consumer goods during the 1970s and 1980s to exporting office and telecommunications equipment in the 1990s. In 1960, agriculture and manufacturing accounted for 59.3 percent and 38 percent of Taiwan's exports, respectively. By 2003, the corresponding figures were 2.5 percent and 93.7 percent. The share of machinery and transport equipment (a subcategory of manufactures) had increased from 1.5 percent in 1960 to 55.7 percent in 2003.

Solution to 2: The challenges of foreign competition created a virtuous circle that was self-reinforcing. Success in export markets increased the confidence of Taiwanese firms and led to greater success in exports through increased productivity, higher-quality products, acquisition of new skills, and adoption of technologies.

Solution to 3: The factors that helped change Taiwan's comparative advantage included government policy, an increasingly skilled and productive workforce, and proactive firms that learned and adapted new technology.

*Prior to the 1960s, Taiwan had an import-substitution policy—that is, a development policy aimed at replacing imports with domestic production that was supported by U.S. aid. However, U.S. aid ended in the 1960s, forcing Taiwan toward export-promotion policies.

From an investment perspective, it is critical for analysts to be able to examine a country's comparative and absolute advantages and to analyze changes in them. It is also important to understand changes in government policy and regulations, demographics, human capital, demand conditions, and other factors that may influence comparative advantage and production and trade patterns. This information can then be used to identify sectors, industries within those sectors, and companies within those industries that will benefit.

2.4.2. Ricardian and Heckscher–Ohlin Models of Comparative Advantage

A discussion of absolute and comparative advantage and the gains from specialization would be incomplete without a discussion of two important theories of trade, the Ricardian model and the Heckscher–Ohlin model. These models are based on cross-country differences in technology and in factor endowments, respectively. These theoretical models are based on several assumptions, some of which may not be fully satisfied in the real world; nonetheless they provide extremely useful insights into the determinants and patterns of trade.

Adam Smith argued that a country could gain from trade if it had an absolute advantage in the production of a good. David Ricardo extended Smith's idea of the gains from trade by arguing that even if a country did not have an absolute advantage in the production of any good, it could still gain from trade if it had a comparative advantage in the production of a good. In the Ricardian model, labor is the only (variable) factor of production. Differences in labor productivity, reflecting underlying differences in technology, are the source of comparative advantage and hence the key driver of trade in this model. A country with a lower opportunity cost in the production of a good has a comparative advantage in that good and will specialize in its production. In our two-country model, if countries vary in size, the smaller country may specialize completely, but may not be able to meet the total demand for the product. Hence, the larger country may be incompletely specialized, producing and exporting the good in which it has a comparative advantage but still producing (and consuming) some of the good in which it has a comparative disadvantage. It is important to recognize that although differences in technology may be a major source of comparative advantage at a given point in time, other countries can close the technology gap or even gain a technological advantage. The shift of information technology (IT) services from developed countries to India is an example of comparative advantage shifting over time.[10] This shift was facilitated by India's growing pool of highly skilled and relatively low-wage labor, the development and growth of its telecommunication infrastructure, and government policies that liberalized trade in the 1990s.

In the Heckscher–Ohlin model (also known as the factor-proportions theory), both capital and labor are variable factors of production. That is, each good can be produced with varying combinations of labor and capital. According to this model, differences in the relative endowment of these factors are the source of a country's comparative advantage. This model assumes that technology in each industry is the same among countries, but it varies among industries. According to the theory, a country has a comparative advantage in goods whose production is intensive in the factor with which it is relatively abundantly endowed, and would tend to specialize in and export that good. Capital is more abundant in a country if the ratio of its endowment of capital to labor is greater than that of its trading partner, and less

[10]According to NASSCOM (India's prominent IT-BPO trade association), Indian firms offer a wide range of information technology services that include consulting, systems integration, IT outsourcing/managed services/hosting services, training, and support/maintenance. See www.nasscom.in.

abundant where the ratio is less.[11] This scenario means a country in which labor is relatively abundant would export relatively labor-intensive goods and import relatively capital-intensive goods. For example, because the manufacture of textiles and clothing is relatively labor intensive, they are exported by such countries as China and India where labor is relatively abundant.

Relative factor intensities in production can be illustrated with the following example. In 2002, capital per worker in the Canadian paper industry was C$118,777, whereas in the clothing manufacturing sector it was C$8,954.[12] These amounts indicate that manufacturing paper is more capital intensive than clothing production. Canada trades with Thailand and, being relatively capital abundant compared with Thailand, it exports relatively capital-intensive paper to Thailand and imports relatively labor-intensive clothing from Thailand.

Because the Heckscher–Ohlin model has two factors of production, labor and capital (unlike the Ricardian model that has only labor), it allows for the possibility of income redistribution through trade. The demand for an input is referred to as a *derived demand* because it is derived from the demand for the product it is used to produce. As a country opens up to trade, it has a favorable impact on the abundant factor and a negative impact on the scarce factor. This is because trade causes output prices to change; more specifically, the price of the exported good increases and the price of the imported good declines. These price changes affect the demand for factors used to produce the imported and exported goods, and hence affect the incomes received by each factor of production.

To illustrate this point, consider again the opening of trade between the United Kingdom and India in Exhibit 8-10a and Exhibit 8-10b. When trade opened, the United Kingdom expanded production of machines—which are assumed to be the capital-intensive industry—and reduced production of clothing. India did the opposite. Machines became more expensive relative to clothing in the United Kingdom (line P^* is steeper than line P^A). The relative price change, along with the shift in output it induces, leads to a redistribution of income from labor to capital in the United Kingdom. The opposite occurs in India—machines become cheaper relative to clothing (line P^* is flatter than line P_A^*), production shifts toward clothing, and income is redistributed from capital to labor.

Note that in each country, the relatively cheap good and the relatively cheap factor of production both get more expensive when trade is opened. That raises an interesting question: If free trade equalizes the prices of goods among countries, does it also equalize the prices of the factors of production? In the simple Heckscher–Ohlin world of homogeneous products, homogeneous inputs, and identical technologies among countries, the answer is yes: The absolute and relative factor prices are equalized in both countries if there is free trade. In the real world, we see that factor prices do not converge completely even if there is free trade, because several assumptions of the models are not fully satisfied in the real world. Nonetheless, it is important to note that *with international trade, factor prices display a tendency to move closer together in the long run.*

Changes in factor endowments can cause changes in the patterns of trade and can create profitable investment opportunities. For example, in 1967 Japan had a comparative advantage in unskilled-labor-intensive goods, such as textiles, apparel, and leather. Meier (1998) notes

[11]Alternatively, factor abundance can be defined in terms of the relative factor prices that prevail in autarky. Under this definition, labor is more abundant in a country if the cost of labor relative to the cost of capital is lower in that country, and less abundant if the cost of labor is higher.

[12]Appleyard, Field, and Cobb (2010, 131).

that by 1980, Japan had greatly increased its skilled labor and consequently had a comparative advantage in skill-intensive products, especially nonelectrical machinery.

It is important to note that technological differences, as emphasized in the Ricardian trade model, and differences in factor abundance, as emphasized in the Heckscher–Ohlin model, are both important drivers of trade. They are complementary, not mutually exclusive. Tastes and preferences can also vary among countries and can change over time, leading to changes in trade patterns and trade flows.

3. TRADE AND CAPITAL FLOWS: RESTRICTIONS AND AGREEMENTS

Trade restrictions (or trade protection) are government policies that limit the ability of domestic households and firms to trade freely with other countries. Examples of trade restrictions include tariffs, import quotas, voluntary export restraints (VERs), subsidies, embargoes, and domestic content requirements. Tariffs are taxes that a government levies on imported goods. Quotas restrict the quantity of a good that can be imported into a country, generally for a specified period of time. A voluntary export restraint is similar to a quota but is imposed by the exporting country. An **export subsidy** is paid by the government to the firm when it exports a unit of a good that is being subsidized. The goal here is to promote exports, but it reduces welfare by encouraging production and trade that are inconsistent with comparative advantage. **Domestic content provisions** stipulate that some percentage of the value added or components used in production should be of domestic origin. Trade restrictions are imposed by countries for several reasons, including protecting established domestic industries from foreign competition, protecting new industries from foreign competition until they mature (infant industry argument), protecting and increasing domestic employment, protecting strategic industries for national security reasons, generating revenues from tariffs (especially for developing countries), and retaliation against trade restrictions imposed by other countries.

Capital restrictions are defined as controls placed on foreigners' ability to own domestic assets and/or domestic residents' ability to own foreign assets. Thus, in contrast with trade restrictions, which limit the openness of goods markets, capital restrictions limit the openness of financial markets. Sections 3.1 through 3.4 discuss trade restrictions. Section 3.5 briefly addresses capital restrictions.

3.1. Tariffs

Tariffs are taxes that a government levies on imported goods.[13] The primary objective of tariffs is to protect domestic industries that produce the same or similar goods. They may also aim to reduce a trade deficit. Tariffs reduce the demand for imported goods by increasing their price above the free trade price. The economic impact of a tariff on imports in a small country is illustrated in Exhibit 8-12. In this context, a small country is not necessarily small in size, population, or GDP. Instead, a small country in this context is one that is a price taker in the world market for a product and cannot influence the world market price. For example, by many measures Brazil is a large country, but it is a price taker in the world market for cars.

[13]Governments may also impose taxes on exports, although they are less common.

In this context, a large country, however, is a large importer of the product and can exercise some influence on price in the world market. When a large country imposes a tariff, the exporter reduces the price of the good to retain some of the market share it could lose if it did not lower its price. This reduction in price alters the terms of trade and represents a redistribution of income from the exporting country to the importing country. So, in theory it is possible for a large country to increase its welfare by imposing a tariff if (1) its trading partner does not retaliate and (2) the deadweight loss as a result of the tariff (see the following discussion) is smaller than the benefit of improving its terms of trade. However, there would still be a net reduction in global welfare—the large country cannot gain by imposing a tariff unless it imposes an even larger loss on its trading partner.

In Exhibit 8-12, the world price (free trade price) is P^*. Under free trade, domestic supply is Q^1, domestic consumption is Q^4, and imports are Q^1Q^4. After the imposition of a per-unit tariff t, the domestic price increases to P_t, which is the sum of the world price and the per-unit tariff t. At the new domestic price, domestic production increases to Q^2 and domestic consumption declines to Q^3, resulting in a reduction in imports to Q^2Q^3.

The welfare effects can be summarized as follows:

- Consumers suffer a loss of consumer surplus because of the increase in price.[14] This effect is represented by areas A + B + C + D in Exhibit 8-12.
- Local producers gain producer surplus from a higher price for their output. This effect is represented by area A.
- The government gains tariff revenue on imports Q^2Q^3. This effect is represented by area C.

The net welfare effect is the sum of these three effects. The loss in consumer surplus is greater than the sum of the gain in producer surplus and government revenue and results in a deadweight loss to the country's welfare of B + D.

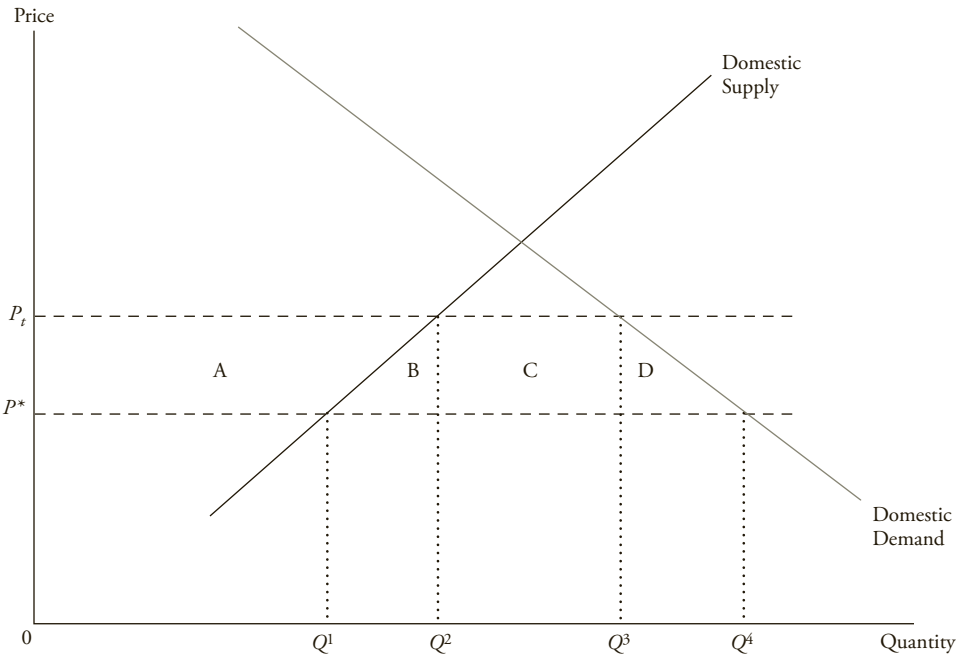
Welfare Effects of an Import Tariff or Quota

	Importing Country
Consumer surplus	−(A + B + C + D)
Producer surplus	+A
Tariff revenue or quota rents	+C
National welfare	$-B - D$

Tariffs create deadweight loss because they give rise to inefficiencies on both the consumption side and the production side. B represents inefficiencies in production. Instead of being able to import goods at the world price P^*, tariffs encourage inefficient producers whose cost of production is greater than P^* to enter (or remain in) the market, leading to an inefficient allocation of resources. On the consumption side, tariffs prevent mutually beneficial exchanges from occurring, because consumers who were willing to pay more than P^* but less than P_t are now unable to consume the good.

[14]Consumer surplus, producer surplus, and deadweight loss are defined and discussed in Chapter 1, "Demand and Supply Analysis: Introduction."

EXHIBIT 8-12 Welfare Effects of Tariff and Import Quota



EXAMPLE 8-5 Analysis of a Tariff

South Africa manufactures 110,000 tons of paper. However, domestic demand for paper is 200,000 tons. The world price for paper is $5 per ton. South Africa will import 90,000 tons of paper from the world market at free trade prices. If the South African government (a small country) decides to impose a tariff of 20 percent on paper imports, the price of imported paper will increase to $6. Domestic production after the imposition of the tariff increases to 130,000 tons, while the quantity demanded declines to 170,000 tons.

1. Calculate the loss in consumer surplus arising from the imposition of the tariff.
2. Calculate the gain in producer surplus arising from the imposition of the tariff.
3. Calculate the change in government revenue arising from the imposition of the tariff.
4. Calculate the deadweight loss arising from the imposition of the tariff.

Solution to 1: The loss in consumer surplus = $1 × 170,000 + ½ × $1 × 30,000 = $185,000. This calculation is represented by areas A + B + C + D in Exhibit 8-12.

Solution to 2: Gain in producer surplus = \$1 × 110,000 + ½ × (\$1 × 20,000) = \$120,000; area A in Exhibit 8-12.

Solution to 3: Change in government revenue = \$1 × 40,000 = \$40,000; area C in Exhibit 8-12.

Solution to 4: Deadweight loss because of the tariff = ½ × \$1 × 20,000 + ½ × \$1 × 30,000 = \$25,000; areas B + D in Exhibit 8-12.

3.2. Quotas

A **quota** restricts the quantity of a good that can be imported into a country, generally for a specified period of time. An import license specifies the quantity that can be imported. For example, the European Union operates a system of annual import quotas for steel producers that are not members of the World Trade Organization. The 2010 quota was 0.2 million tons a year for Kazakhstan. In the case of Russia, the 2010 quota of 3.2 million tons per year was a part of an EU–Russia agreement.[15] A key difference between tariffs and quotas is that the government is able to collect the revenue generated from a tariff. This effect is uncertain under a quota. With quotas, foreign producers can often raise the prices of their goods and earn greater profits than they would without the quota. These profits are called **quota rents**. In Exhibit 8-12, if the quota is Q^2Q^3, the equivalent tariff that will restrict imports to Q^2Q^3 is t and the domestic price after the quota is P_t. This is the same as the domestic price after the tariff t was imposed. Area C, however, is now the quota rent or profits that are likely to be captured by the foreign producer rather than tariff revenue that is captured by the domestic government. If the foreign producer or foreign government captures the quota rent, C, then the welfare loss to the importing country, represented by areas B + D + C in Exhibit 8-12, under a quota is greater than under the equivalent tariff. If the government of the country that imposes the quota can capture the quota rents by auctioning the import licenses for a fee, then the welfare loss under the quota is similar to that of a tariff, represented by areas B + D.

A **voluntary export restraint** (VER) is a trade barrier under which the exporting country agrees to limit its exports of the good to its trading partners to a specific number of units. The main difference between an import quota and a VER is that the former is imposed by the importer, whereas the latter is imposed by the exporter. The VER allows the quota rent resulting from the decrease in trade to be captured by the exporter (or exporting country), whereas in the case of an import quota there is ambiguity regarding who captures the quota rents. Hence, a VER results in welfare loss in the importing country. For example, in 1981 the Japanese government imposed VERs on automobile exports to the United States.

3.3. Export Subsidies

An export subsidy is a payment by the government to a firm for each unit of a good that is exported. Its goal is to stimulate exports. But it interferes with the functioning of the free market and may distort trade away from comparative advantage. Hence, it reduces welfare. *Countervailing duties* are duties that are levied by the importing country against subsidized

[15]For more information, see http://ec.europa.eu/trade/creating-opportunities/economic-sectors/industrial-goods/steel/.

exports entering the country. As an example, agricultural subsidies in developed countries, notably the EU, have been a contentious issue in trade negotiations with less developed countries and developed countries that are agricultural exporters, such as New Zealand and Australia.

In the case of an export subsidy, the exporter has the incentive to shift sales from the domestic to the export market because it receives the international price plus the per-unit subsidy for each unit of the good exported. This scenario raises the price in the domestic market by the amount of the subsidy in the small country case (price before subsidy plus subsidy). In the large country case, the world price declines as the large country increases exports. The net welfare effect is negative in both the large and small country cases, with a larger decline in the large country case. This result is because in the large country case, the decline in world prices implies that a part of the subsidy is transferred to the foreign country, unlike in the small country case.

Exhibit 8-13 summarizes some of these effects.

EXHIBIT 8-13 Effects of Alternative Trade Policies

	Tariff	Import Quota	Export Subsidy	VER
Impact on	Importing country	Importing country	Exporting country	Importing country
Producer surplus	Increases	Increases	Increases	Increases
Consumer surplus	Decreases	Decreases	Decreases	Decreases
Government revenue	Increases	Mixed (depends on whether the quota rents are captured by the importing country through sale of licenses or by the exporters)	Falls (government spending rises)	No change (rent to foreigners)
National welfare	Decreases in small country Could increase in large country	Decreases in small country Could increase in large country	Decreases	Decreases

Effects on Price, Production, Consumption, and Trade

	Tariff	Import Quota	Export Subsidy	VER
Impact on	Importing country	Importing country	Exporting country	Importing country
Price	Increases	Increases	Increases	Increases
Domestic production	Increases	Increases	Increases	Increases
Domestic consumption	Decreases	Decreases	Decreases	Decreases
Trade	Imports decrease	Imports decrease	Exports increase	Imports decrease

EXAMPLE 8-6 Tariffs, Quotas, and VERs

Thailand, a small country, has to decide whether to impose a tariff or a quota on the import of computers. You are considering investing in a local firm that is a major importer of computers.

1. What will be the impact of a tariff on prices, quantity produced, and quantity imported in Thailand (the importing country)?
2. If Thailand imposes a tariff, what will the impact be on prices in the exporting country?
3. How would a tariff affect consumer surplus, producer surplus, and government revenue in Thailand?
4. Explain whether the net welfare effect of a tariff is the same as that of a quota.
5. Which policy, a tariff or a quota, would be more beneficial to the local importer in which you may invest, and why?
6. If Thailand were to negotiate a VER with the countries from which it imports computers, would this be better or worse than an import quota for the local importing firm in which you may invest? Why?

Solution to 1: A tariff imposed by a small country, such as Thailand, raises the prices of computers in the importing country, increases domestic production, and reduces the quantity imported.

Solution to 2: A tariff imposed by a small country would not change the prices of computers in the exporting country.

Solution to 3: When a small country imposes a tariff, it reduces consumer surplus, increases producer surplus, and increases government revenue in that country.

Solution to 4: The quota can lead to a greater welfare loss than a tariff if the quota rents are captured by the foreign government or foreign firms.

Solution to 5: A tariff will hurt importers because it will reduce their share of the computer market in Thailand. The impact of a quota depends on whether the importers can capture a share of the quota rents. Assuming importers can capture at least part of the rents, they will be better off with a quota.

Solution to 6: The VER would not be better for the local importer than the import quota and would most likely be worse. Under the VER, all of the quota rents will be captured by the exporting countries, whereas with an import quota at least part of the quota rents may be captured by local importers.

It is important to understand existing trade policies and the potential for policy changes that may impact return on investment. Changes in the government's trade policy can affect the pattern and value of trade and may result in changes in industry structure. These changes may have important implications for firm profitability and growth because they can affect the goods a firm can import and export, change demand for its products, impact its pricing policies, and create delays through increased paperwork, procurement of licenses, approvals,

and so on. For example, changes in import policies that affect the ability of a firm to import vital inputs for production may increase the cost of production and reduce firm profitability.

3.4. Trading Blocs, Common Markets, and Economic Unions

There has been a proliferation of trading blocs or regional trading agreements (RTAs) in recent years. Important examples of regional integration include the North American Free Trade Agreement (NAFTA) and the European Union (EU). A regional trading bloc is a group of countries that have signed an agreement to reduce and progressively eliminate barriers to trade and movement of factors of production among the members of the bloc. The agreement may or may not have common trade barriers against countries that are not members of the bloc.

There are many different types of regional trading blocs, depending on the level of integration that takes place. **Free trade areas** (FTAs) are one of the most prevalent forms of regional integration. In FTAs all barriers to the flow of goods and services among members have been eliminated. However, each country maintains its own polices against nonmembers. The North American Free Trade Agreement among the United States, Canada, and Mexico is an example of an FTA. A **customs union** extends the FTA by not only allowing free movement of goods and services among members but also creating a common trade policy against nonmembers. In 1947, Belgium, the Netherlands, and Luxemburg ("Benelux") formed a customs union that became a part of the European Community in 1958. The **common market**, the next level of economic integration, incorporates all aspects of the customs union and extends it by allowing free movement of factors of production among members. The Southern Cone Common Market (MERCOSUR) of Argentina, Brazil, Paraguay, and Uruguay is an example of a common market.[16] An **economic union** requires an even greater degree of integration. It incorporates all aspects of a common market and in addition requires common economic institutions and coordination of economic policies among members. The European Community became the European Union in 1993. If the members of the economic union decide to adopt a common currency, then it is also a **monetary union**. For example, with the adoption of the euro, 11 EU member countries also formed a monetary union.[17]

Regional integration is popular because eliminating trade and investment barriers among a small group of countries is easier, politically less contentious, and quicker than multilateral trade negotiations under the World Trade Organization (WTO). The WTO is a negotiating forum that deals with the rules of global trade between nations and is where member countries can go to sort out trade disputes. The latest rounds of trade negotiations launched by the WTO in 2001 at Doha, Qatar, included several contentious issues of specific concern to developing countries, such as the cost of implementing trade policy reform in such countries, market access in developed countries for developing countries' agricultural products, and

[16]For more information, visit the OECD website, http://stats.oecd.org/glossary/.

[17]On 1 January 1999, Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain adopted the euro. This adoption meant that these countries had to surrender control over their domestic monetary policy to the European Central Bank. Greece joined in 2001. Euro coins and notes went into circulation on 1 January 2002, and these countries gave up the last vestiges of their national currencies. Other members now include Slovenia (2007), Cyprus (2008), Malta (2008), Slovakia (2009), and Estonia (2011). The eurozone (i.e., the monetary union) is only a subset of the EU membership because some EU members, notably the United Kingdom, have not adopted the euro.

EXAMPLE 8-7 Trading Blocs

1. Chile and Australia have free trade with each other but have separate trade barriers on imports from other countries. Chile and Australia are part of a(n):
   A. FTA.
   B. economic union.
   C. customs union.
   D. common market.
2. An RTA that removes all tariffs on imports from member countries and has common external tariffs against all nonmembers but does not advance further in deepening economic integration is called a(n):
   A. FTA.
   B. economic union.
   C. customs union.
   D. common market.

Solution to 1: A is correct. Chile and Australia do not a have customs union because they do not have a common trade policy with respect to other trade partners (C is incorrect). An economic union and a common market entail even more integration (B and D are incorrect).

Solution to 2: C is correct. A basic FTA does not entail common external tariffs (A is incorrect), whereas an economic union and a common market entail integration beyond common external tariffs (B and D are incorrect).

access to affordable pharmaceuticals in developing countries. After nearly a decade of negotiations, very limited progress has been made on the major issues. Hence, it is not surprising to see a renewed interest in bilateral and multilateral trade liberalization on a smaller scale. Policy coordination and harmonization are also easier among a smaller group of countries. Regional integration can be viewed as a movement toward freer trade.

Regional integration results in preferential treatment for members compared with nonmembers and can lead to changes in the patterns of trade. Member countries move toward freer trade by eliminating or reducing trade barriers against each other, leading to a more efficient allocation of resources. But regional integration may also result in trade and production being shifted from a lower-cost nonmember that still faces trade barriers to a higher-cost member who faces no trade barriers. This shift leads to a less efficient allocation of resources and could reduce welfare. Hence, there are two static effects that are direct results of the formation of the customs union: trade creation and trade diversion.

Trade creation occurs when regional integration results in the replacement of higher-cost domestic production by lower-cost imports from other members. For example, consider two hypothetical countries, Qualor and Vulcan. Qualor produces 10 million shirts annually and imports 2 million shirts from Vulcan, which has a lower cost of production. Qualor has 10 percent tariffs on imports from Vulcan. Qualor and Vulcan then agree to form a customs union. Qualor reduces its production of shirts to 7 million and now imports 11 million shirts

from Vulcan. The decline in Qualor's domestic production (from 10 million to 7 million shirts) is replaced by importing 3 million additional shirts from the low-cost producer, Vulcan; this scenario represents trade creation. The rest of the additional imports (6 million shirts) represent increased consumption by Qualor's consumers because the price of shirts declines after formation of the custom union.

Trade diversion occurs when lower-cost imports from nonmember countries are replaced with higher-cost imports from members. In the example in the preceding paragraph, suppose Qualor initially imposes a 10 percent tariff on imports from both Vulcan and Aurelia. Aurelia is the lowest-cost producer of shirts, so Qualor initially imports 2 million shirts from Aurelia instead of from Vulcan. Qualor and Vulcan then form a customs union, which eliminates tariffs on imports from Vulcan but maintains a 10 percent tariff on imports from Aurelia. Now trade diversion could occur if the free trade price on imports from Vulcan is lower than the price on imports from Aurelia. Even though Aurelia is the lowest-cost producer, it may be a higher-priced source of imports because of the tariff. If this is the case, then Qualor will stop importing from Aurelia, a nonmember, and divert its imports to Vulcan, a member of the RTA. Both trade creation and trade diversion are possible in an RTA. If trade creation is larger than trade diversion, then the net welfare effect is positive. However, there are concerns that this may not always be the case.

The benefits ascribed to free trade—greater specialization according to comparative advantage, reduction in monopoly power because of foreign competition, economies of scale from larger market size, learning by doing, technology transfer, knowledge spillovers, greater foreign investment, and better-quality intermediate inputs at world prices—also apply to regional trading blocs. In addition, fostering greater interdependence among members of the regional trading bloc reduces the potential for conflict. Members of the bloc also have greater bargaining power and political clout in the global economy by acting together instead of as individual countries.

The 2009 *World Development Report* (World Bank 2009) points to spillover of growth across borders as one of the main benefits of regional integration (Collier and O'Connell 2007). There is evidence of considerable spillovers among OECD countries, which are highly integrated both as a group and within their own geographic regions. The long-run growth of integrated countries is interconnected because members have greater access to each other's markets. Strong growth in any RTA country could have a positive impact on growth in other RTA member countries. RTAs also enhance the benefits of good policy and lead to convergence in living standards. For example, growth spillovers are likely to be much smaller among sub-Saharan African countries because of a lack of integration arising from deficiencies in RTAs and inadequate levels of transportation and telecommunications infrastructure. Roberts and Deichmann (2008) estimated what the cumulative loss in real GDP between 1970 and 2000 would have been if Switzerland, which is landlocked and fully integrated with both its immediate neighbors and the world economy, had been subject to the same level of spillovers as the Central African Republic. Under such a scenario, Switzerland's GDP per capita in 2000 would have been 9.3 percent lower. The cumulative GDP loss would have been $334 billion (constant U.S. dollars, 2000), which was the equivalent of 162 percent of Switzerland's real GDP in 2000.

Although regional integration has many advantages, it may impose costs on some groups. For example, there was significant concern in the United States that NAFTA and especially low-skilled-labor-intensive imports from Mexico could hurt low-skilled U.S. workers. Adjustment costs arose as import competition caused inefficient U.S. firms to exit the market, and the workers in those firms were at least temporarily unemployed as they sought new jobs.

However, the surviving firms experienced an increase in productivity, and U.S. consumers benefited from the increase in product varieties imported from Mexico. Feenstra and Taylor (2008) estimated that the product varieties exported from Mexico to the United States had grown by an average of 2.2 percent a year across all industries. They estimated that NAFTA imposed private costs of nearly $5.4 billion a year in the United States during 1994–2002, but that these costs were offset by an average welfare gain of $5.5 billion a year accruing from increased varieties imported from Mexico. Consumer gains from more varieties of products continued over time as long as the imports continued, while adjustment costs arising from job losses declined over time. In 2003, the gain from increased product varieties from Mexico was $11 billion, far exceeding the adjustment costs of $5.4 billion.[18] Their analysis concluded:

> Thus the consumer gains from increased product variety, when summed over the years, considerably exceed the private loss from displacement. This outcome is guaranteed to occur because the gains from expanded import varieties occur every year that the imports are available, whereas labor displacement is a temporary phenomenon. (Feenstra and Taylor 2008, 208)

It is important to recognize, however, that workers displaced by regional integration may have to bear long-term losses if they are unable to find jobs with wages comparable with the jobs they lost or they remain unemployed for a long period. For example, although import competition was certainly not the only factor that led to a dramatic contraction of the U.S. automobile industry, the impact on employment in that industry is likely to be permanent, and many former autoworkers, especially older workers, may never find comparable jobs.

Concerns regarding national sovereignty, especially where big and small nations may be part of the same bloc, have also been an impediment to the formation of FTAs. The proposal for a South Asian regional bloc has faced challenges regarding India's role because it is one of the biggest economies in the region.

Regional integration is important from an investment perspective because it offers new opportunities for trade and investment. The cost of doing business in a large, single, regional market is lower, and firms can benefit from economies of scale. However, it is important to note that differences in tastes, culture, and competitive conditions still exist among members of a trading bloc. These differences may limit the potential benefits from investments within the bloc. In addition, depending on the level of integration and the safeguards in place, problems faced by individual member countries in an RTA may quickly spread to other countries in the bloc.

There are at least two challenges in the formation of an RTA and in its potential progression from a free trade area to deeper integration in the form of a customs union, common market, or economic union. First, cultural differences and historical considerations—for example, wars and conflicts—may complicate the social and political process of integration. Second, maintaining a high degree of economic integration limits the extent to which member countries can pursue independent economic and social policies. Free trade and mobility of labor and capital tend to thwart policies aimed at controlling relative prices and quantities within a country, while balance of payments and fiscal credibility considerations limit the viability of divergent macroeconomic policies. This situation is especially true

[18]Feenstra and Taylor (2008) discuss in their book on pages 207–208 the data limitations and various assumptions they made in their analysis.

in the case of a monetary union because monetary policy is not under the control of individual countries, and currency devaluation and revaluation are not available as a tool to correct persistent imbalances.[19] When persistent imbalances do arise, they may lead to a crisis that spills over to other countries facing similar problems. A recent example is the fear of contagion caused by the Greek fiscal crisis in 2010. In May 2010, Standard & Poor's reduced the credit ratings on Greece's government from investment grade to junk status. It also downgraded the government debt of Spain and Portugal. These countries were suffering from a combination of high government deficits and slow GDP growth. The credit downgrades increased fears that Greece, in particular, would default on its debt and cause economic turmoil not only among the healthier countries in the EU but also in the United States and Asia. The EU and the International Monetary Fund (IMF) agreed on a USD145 billion (EUR110 billion) bailout for Greece in May 2010, and provided Ireland with a financing package of about USD113 billion (EUR85 billion) in November 2010. As of late 2010, there were continuing concerns about the financial health of Greece, Ireland, Portugal, and Spain. The EU, which created the European Financial Stability Facility (EFSF) in 2010 to help EU countries in need, has been debating the need for an expansion in the scope and financing capacity of the EFSF.

EXAMPLE 8-8 Trade Agreements

Bagopia, Cropland, and Technopia decide to enter into an RTA. In the first stage, they sign an agreement to form a free trade area (FTA). After several successful years, they decide that it is time to form a common market.

1. Does an FTA make exporting firms in member countries more attractive as investment options?
2. How does the common market affect firms doing business in these countries compared with an FTA?

Solution to 1: The first stage, in which there is free movement of goods and services among RTA members, is called a free trade area. It makes exporting firms a more attractive investment proposition because they are able to serve markets in member countries without the additional costs imposed by trade barriers.

Solution to 2: Unlike an FTA, a common market allows for free movement of factors of production, such as labor and capital, among the member economies. Like an FTA, it provides access to a much larger market and free movement of goods and services. But the common market can create more profitable opportunities for firms than an FTA by allowing them to locate production in and purchase components from anywhere in the common market according to comparative advantage.

[19]These limitations are inherent in any system with fixed exchange rates and a high degree of capital mobility. They are not unique to a monetary union (i.e., a common currency). For a discussion of currency regimes, see Chapter 9, "Currency Exchange Rates."

3.5. Capital Restrictions

There are many reasons for governments to restrict inward and outward flow of capital. For example, the government may want to meet some objective regarding employment or regional development, or it may have a strategic or defense-related objective. Many countries require approval for foreigners to invest in the country and for citizens to invest abroad. Control over inward investment by foreigners results in restrictions on how much can be invested, and on the type of industries in which capital can be invested. For example, such strategic industries as defense and telecommunications are often subject to ownership restrictions. Outflow restrictions can include restrictions on repatriation of capital, interest, profits, royalty payments, and license fees. Citizens are often limited in their ability to invest abroad, especially in foreign-exchange-scarce economies, and there can be deadlines for repatriation of income earned from any investments abroad.

Economists consider free movement of financial capital to be beneficial because it allows capital to be invested where it will earn the highest return. Inflows of capital also allow countries to invest in productive capacity at a rate that is higher than could be achieved with domestic savings alone, and it can enable countries to achieve a higher rate of growth. Longer-term investments by foreign firms that establish a presence in the local economy can bring in not only much needed capital but also new technology, skills, and advanced production and management practices, as well as create spillover benefits for local firms. Investment by foreign firms can also create a network of local suppliers if they source some of their components locally. Such suppliers may receive advanced training and spillover benefits from close working relationships with the foreign firms. On the one hand, increased competition from foreign firms in the market may force domestic firms to become more efficient. On the other hand, it is possible that the domestic industry may be hurt because domestic firms that are unable to compete are forced to exit the market.

In times of macroeconomic crisis, capital mobility can result in capital flight out of the country, especially if most of the inflow reflects short-term portfolio flows into stocks, bonds, and other liquid assets rather than foreign direct investment (FDI) in productive assets. In such circumstances, capital restrictions are often used in conjunction with other policy instruments, such as fixed exchange rate targets. Capital restrictions and fixed exchange rate targets are complementary instruments because in a regime of perfect capital mobility, governments cannot achieve domestic and external policy objectives simultaneously using only standard monetary and fiscal policy tools.[20] By limiting the free flow of capital, capital controls provide a way to exercise control over a country's external balance whereas more traditional macroeconomic policy tools are used to address other objectives. As an example, China has pegged its currency to the U.S. dollar in a narrow range. At the same time, it limits the free flow of capital into and out of the country. The capital controls serve two purposes. First, they make it easier to maintain the tight exchange rate peg that is crucial in fostering the nation's export sector. Second, the capital controls shield domestic interest rates from external market forces. Control over domestic interest rates is crucial for managing the domestic banking and real estate sectors. In essence, the capital controls allow China to exercise a degree of monetary policy independence that would not be achievable under a fixed exchange rate regime with free capital flows.

[20]Section 4.1 of Chapter 9, "Currency Exchange Rates," provides a concise discussion of the policy implications of capital mobility with fixed versus floating exchange rates.

Modern capital controls were developed by the belligerents in World War I as a method to finance the war effort. At the start of the war, all major powers restricted capital outflows (i.e., the purchase of foreign assets or loans abroad). These restrictions raised revenues by keeping capital in the domestic economy, facilitating the taxation of wealth, and producing interest income. Moreover, capital controls helped to maintain a low level of interest rates, reducing the government's borrowing costs on its liabilities. Since World War I, controls on capital outflows have been used similarly in other countries, mostly developing nations, to generate revenue for governments or to permit them to allocate credit in the domestic economy without risking capital flight. In broad terms, a capital restriction is any policy designed to limit or redirect capital flows. Such restrictions may take the form of taxes, price or quantity controls, or outright prohibitions on international trade in assets. Price controls may take the form of special taxes on returns to international investment, taxes on certain types of transactions, or mandatory reserve requirements—that is, requirements forcing foreign parties wishing to deposit money in a domestic bank account to deposit some percentage of the inflow with the central bank for a minimum period at zero interest. Quantity restrictions on capital flows may include rules imposing ceilings or requiring special authorization for new or existing borrowing from foreign creditors. Or there may be administrative controls on cross-border capital movements in which a government agency must approve transactions for certain types of assets.

Effective implementation of capital restrictions may entail nontrivial administration costs, particularly if the measures have to be broadened to close potential loopholes. There is also the risk that protecting the domestic financial markets by capital restrictions may postpone necessary policy adjustments or impede private-sector adaptation to changing international circumstances. Most importantly, controls may give rise to negative market perceptions, which may, in turn, make it more costly and difficult for the country to access foreign funds.

In a study on the effectiveness of capital controls, the International Monetary Fund considered restrictions on capital outflows and inflows separately.[21] The authors concluded that for restrictions on capital inflows to be effective (i.e., not circumvented) the coverage needs to be comprehensive and the controls need to be implemented forcefully. Considerable administrative costs are incurred in continuously extending, amending, and monitoring compliance with the regulations. Although controls on inflows appeared to be effective in some countries, it was difficult to distinguish the impact of the controls from the impact of other policies, such as strengthening of prudential regulations, increased exchange rate flexibility, and adjustment of monetary policy. In the case of capital outflows, the imposition of controls during episodes of financial crisis seems to have produced mixed results, providing only temporary relief of varying duration to some countries, while successfully shielding others (e.g., Malaysia) and providing them with sufficient time to restructure their economies.

4. THE BALANCE OF PAYMENTS

The **balance of payments** (BOP) is a double-entry bookkeeping system that summarizes a country's economic transactions with the rest of the world for a particular period of time, typically a calendar quarter or year. In this context, a transaction is defined as "an economic

[21]Ariyoshi et al. (2000).

EXAMPLE 8-9 Capital Restrictions: Malaysia's Capital Controls in 1998–2001

After the devaluation of the Thai baht in July 1997, Southeast Asia suffered from significant capital outflows that led to falling local equity and real estate prices and declining exchange rates. To counter the outflows of capital, the IMF urged many of the countries in the region to increase interest rates, thus making their assets more attractive to foreign investors. Higher interest rates, however, weighed heavily on the domestic economies. In response to this dilemma, Malaysia imposed capital controls on 1 September 1998. These controls prohibited transfers between domestic and foreign accounts, eliminated credit facilities to offshore parties, prevented repatriation of investment until 1 September 1999, and fixed the exchange rate of the Malaysian ringgit at 3.8 per U.S. dollar. In February 1999, a system of taxes on capital flows replaced the prohibition on repatriation of capital. Although the details were complex, the net effect was to discourage short-term capital flows while permitting long-term transactions. By imposing capital controls, Malaysia hoped to regain monetary independence, and to be able to cut interest rates without provoking a fall in the value of its currency as investors avoided Malaysian assets. The imposition of outflow controls indeed curtailed speculative capital outflows and allowed interest rates to be reduced substantially. At the same time, under the umbrella of the capital controls, the authorities pursued bank and corporate restructuring and achieved a strong economic recovery in 1999 and 2000. With the restoration of economic and financial stability, administrative controls on portfolio outflows were replaced by a two-tier, price-based exit system in February 1999, which was finally eliminated in May 2001. Although Malaysia's capital controls did contribute to a stabilization of its economy, they came with long-term costs associated with the country's removal from the Morgan Stanley Capital International (MSCI) developed equity market index, an important benchmark in the institutional asset management industry, and its relegation to the emerging market universe. The Malaysian market was no longer seen as on par with developed equity markets whose institutional and regulatory frameworks provide a higher standard of safety for investors. As a consequence, it became more difficult for Malaysia to attract net long-term capital inflows (Kawai and Takagi 2003).

1. Under what economic circumstances were Malaysia's capital restrictions imposed?
2. What was the ultimate objective of Malaysia's capital restrictions?
3. How successful were the country's capital restrictions?

Solution to 1: As a result of the Southeast Asian crisis, Malaysia suffered substantial net capital outflows, pushing up the domestic interest rate level.

Solution to 2: The restrictions were designed to limit and redirect capital flows to allow the government to reduce interest rates and pursue bank and corporate restructurings.

Solution to 3: Although the capital controls helped stabilize Malaysia's economy, they contributed to a change in investors' perception of Malaysian financial markets and removal of the Malaysian equity market from the MSCI benchmark universe of developed equity markets. This situation undermined international demand for Malaysian equities and made it more difficult to attract net long-term capital inflows.

flow that reflects the creation, transformation, exchange, or extinction of economic value and involves changes in ownership of goods and/or financial assets, the provision of services, or the provision of labour and capital."[22] In other words, the balance of payments reflects payments for exports and imports as well as financial transactions and financial transfers. Analyzing the BOP is an important element in assessing a country's macroeconomic environment, its monetary and fiscal policies, and its long-term growth potential. Investors use data on trade and capital flows to evaluate a country's overall level of capital investment, profitability, and risk. The following section describes the balance of payments, the factors that influence it, and its impact on exchange rates, interest rates, and capital market transactions.

4.1. Balance of Payments Accounts

The balance of payments is a double-entry system in which every transaction involves both a debit and a credit. In principle, the sum of all debit entries should equal the sum of all credit entries, and the net balance of all entries on the BOP statement should equal zero. In practice, however, this is rarely the case because the data used to record balance of payments transactions are often derived from different sources.

Debit entries reflect purchases of imported goods and services, purchases of foreign financial assets, payments received for exports, and payments (interest and principal) received from debtors. Credit entries reflect payments for imported goods and services, payments for purchased foreign financial assets, and payments to creditors (see Exhibit 8-14, Panel A). Put differently, a debit represents an increase in a country's assets (the purchase of foreign assets or the receipt of cash from foreigners) or a decrease in its liabilities (the amount owed to foreigners); a credit represents a decrease in assets (the sale of goods and services to foreigners or the payment of cash to foreigners) or an increase in liabilities (an amount owed to foreigners).

For example, as shown in Panel B of Exhibit 8-14, on 1 September Country A purchases $1 million of goods from Country B and agrees to pay for these goods on 1 December. On 1 September, Country A would record in its BOP a $1 million debit to reflect the value of the goods purchased (i.e., increase in assets) and $1 million credit to reflect the amount owed to Country B. On 1 December, Country A would record in its BOP a $1 million debit to reflect a decrease in the amount owed (liability) to Country B and a $1 million credit to reflect the actual payment to Country B (decrease in assets).

From Country B's perspective, on 1 September it would record in its BOP a $1 million debit to reflect the amount owed by Country A and a $1 million credit to reflect the sale of goods (exports). On 1 December, Country B would record a $1 million debit to reflect the cash received from Country A and a $1 million credit to reflect the fact that it is no longer owed $1 million by Country A.

4.2. Balance of Payments Components

The BOP is composed of the **current account** that measures the flow of goods and services, the **capital account** that measures transfers of capital, and the **financial account** that records investment flows. These accounts are further disaggregated into subaccounts.

[22]International Monetary Fund (2010a, ch. II, p. 6).

EXHIBIT 8-14 Basic Entries in a Balance of Payments Context

Panel A

Debits Increase in Assets, Decrease in Liabilities	Credits Decrease in Assets, Increase in Liabilities
• Value of imported goods and services • Purchases of foreign financial assets • Receipt of payments from foreigners • Increase in debt owed by foreigners • Payment of debt owed to foreigners	• Payments for imports of goods and services • Payments for foreign financial assets • Value of exported goods and services • Payment of debt by foreigners • Increase in debt owed to foreigners

Panel B

Country A	Debits	Credits
1 September	$1 million Goods purchased from Country B *(increase in real assets)*	$1 million Short-term liability for goods purchased from Country B *(increase in financial liabilities)*
1 December	$1 million Elimination of short-term liability for goods purchased from Country B *(decrease in financial liabilities)*	$1 million Payment for goods purchased from Country B *(decrease in financial assets)*

Country B	Debits	Credits
1 September	$1 million Short-term claim for goods delivered to Country A *(increase in financial assets)*	$1 million Goods delivered to Country A *(decrease in real assets)*
1 December	$1 million Receipt of payment for goods delivered to Country A *(increase in financial assets)*	$1 million Elimination of claim for goods delivered to Country A *(decrease in financial assets)*

4.2.1. Current Account

The current account can be decomposed into four subaccounts:

1. Merchandise trade consists of all commodities and manufactured goods bought, sold, or given away.
2. Services include tourism, transportation, engineering, and business services such as legal services, management consulting, and accounting. Fees from patents and copyrights on new technology, software, books, and movies are also recorded in the services category.
3. Income receipts include income derived from ownership of assets, such as dividends and interest payments; income on foreign investments is included in the current account

because that income is compensation for services provided by foreign investments. When a German company builds a plant in China, for instance, the services the plant generates are viewed as a service export from Germany to China equal in value to the profits the plant yields for its German owner.
4. Unilateral transfers represent one-way transfers of assets, such as worker remittances from abroad to their home countries and foreign direct aid or gifts.

4.2.2. Capital Account

The capital account consists of two subaccounts:

1. Capital transfers include debt forgiveness and migrants' transfers (goods and financial assets belonging to migrants as they leave or enter the country).[23] Capital transfers also include the transfer of title to fixed assets and the transfer of funds linked to the sale or acquisition of fixed assets, gift and inheritance taxes, death duties, uninsured damage to fixed assets, and legacies.
2. Sales and purchases of nonproduced, nonfinancial assets, such as the rights to natural resources and the sale and purchase of intangible assets such as patents, copyrights, trademarks, franchises, and leases.

4.2.3. Financial Account

The financial account can be broken down in two subaccounts: financial assets abroad and foreign-owned financial assets within the reporting country.

1. A country's assets abroad are further divided into official reserve assets, government assets, and private assets. These assets include gold, foreign currencies, foreign securities, the government's reserve position in the International Monetary Fund,[24] direct foreign investment, and claims reported by resident banks.
2. Foreign-owned assets in the reporting country are further divided into official assets and other foreign assets. These assets include securities issued by the reporting country's government and private sectors (e.g., bonds, equities, mortgage-backed securities), direct investment, and foreign liabilities reported by the reporting country's banking sector.

4.3. Paired Transactions in the Balance of Payments Bookkeeping System

The following examples illustrate how some typical cross-border transactions are recorded in the balance of payments framework outlined previously. They include commercial exports and imports, the receipt of income from foreign investments, loans made to borrowers abroad, and purchases of home-country currency by foreign central banks. Exhibit 8-16 illustrates the various individual bookkeeping entries from the perspective of an individual country, in this case Germany.

[23]Immigrants bring with them goods and financial assets already in their possession. Hence, these goods are imported on grounds other than commercial transactions.
[24]These are in effect official currency reserves held with the International Monetary Fund.

EXAMPLE 8-10 U.S. Current Account Balance

Exhibit 8-15 shows a simplified version of the U.S. balance of payments for 1970–2009.

EXHIBIT 8-15 U.S. International Transactions Accounts Data, 1970–2009

	(USD millions)					
(Credits+, Debits−)	1970	1980	1985	1990	2000	2009
Current Account						
Exports of goods and services and income receipts	68,387	344,440	387,612	706,975	1,421,515	2,159,000
Exports of goods and services	56,640	271,834	289,070	535,233	1,070,597	1,570,797
Income receipts	11,748	72,606	98,542	171,742	350,918	588,203
Imports of goods and services and income payments	−59,901	−333,774	−483,769	−759,290	−1,779,241	−2,412,489
Imports of goods and services	−54,386	−291,241	−410,950	−616,097	−1,449,377	−1,945,705
Income payments	−5,515	−42,532	−72,819	−143,192	−329,864	−466,783
Unilateral current transfers, net	−6,156	−8,349	−21,998	−26,654	−58,645	−124,943
Capital Account						
Capital account transactions, net	. . .	. . .	. . .	−7,220	−1	−140
Financial Account						
U.S.-owned assets abroad, ex derivatives (increase/ financial outflow)	−9,337	−86,967	−44,752	−81,234	−560,523	−140,465
Foreign-owned assets in U.S., ex derivatives (increase/financial inflow)	7,226	62,037	144,231	139,357	1,038,224	305,736
Financial derivatives, net	NA	NA	NA	NA	NA	50,804
Statistical discrepancy (sum of above items with sign reversed)	−219	22,613	18,677	28,066	−61,329	162,497

Based only on the information given, address the following:

1. Calculate the current account balance for each year.
2. Calculate the financial account balance for each year.
3. Describe the long-term change in the current account balance.
4. Describe the long-term change in the financial account balance.

Solutions to 1 and 2:

(Credits+, Debits–)	1970	1980	1985	1990	2000	2009
Current account	2,330	2,317	−118,155	−78,969	−416,371	−378,432
Financial account	−2,111	−24,930	99,479	58,123	477,701	216,075

Solution to 3: The United States had a current account surplus until 1980. After 1985, the U.S. current account had a continuing deficit as a result of strong import growth.

Solution to 4: Mirroring the U.S. current account deficit, the U.S. financial account, after 1985, registered continuing net capital inflows in similar proportions to the deficit in the current account.

4.3.1. Commercial Exports: Transactions (ia) and (ib)

A company in Germany sells technology equipment to a South Korean automobile manufacturer for a total price of EUR50 million, including freight charges of EUR1 million to be paid within 90 days. The merchandise will be shipped via a German cargo ship. In this case, Germany is exporting two assets: equipment and transportation services. The cargo shipped is viewed as being created in Germany and used by South Korean customers. In return for relinquishing these two assets, Germany acquires a financial asset—the promise by the South Korean manufacturer to pay for the equipment in 90 days.

Germany would record a EUR50 million debit to an account called "private short-term claims" to show an increase in this asset. It would also record a credit of EUR49 million to "goods" and another credit of EUR1 million to "services." Both credit entries are listed in the exports category and show the decrease in assets available to German residents. These figures are entered as credits on lines 2 and 3 and as a debit on 19 in Exhibit 8-16 and are marked with (ia) to identify a typical commercial export transaction. To pay for the technology equipment purchased from Germany, the South Korean auto manufacturer may purchase euros from its local bank (i.e., a EUR demand deposit held by the South Korean bank in a German bank) and then transfer them to the German exporter. As a result, German liabilities to South Korean residents (i.e., South Korean private short-term claims) would be debited. The respective entries, marked with (ib), are on lines 19 and 23 in Exhibit 8-16.

EXHIBIT 8-16 Hypothetical Transactions between German Residents and Foreigners

Item #	Account	Debit −	Credit +	Balance +/−
1	**Exports of goods and services, income received**			**55**
2	Goods		49 (ia)	49
3	Services		1 (ia)	1
4	Income on residents' investments abroad		5 (v)	5
5	**Imports of goods and services, income paid**			**−45**
6	Goods	45 (ii)		−45
7	Services			
8	Income on foreign investments in home country			
9	**Unilateral transfers**			
10	**Changes in residents' claims on foreigners**			**−105**
11	Official reserve assets			
12	Gold			
13	Foreign currency balances			
14	Other			
15	Government claims			
16	Private claims			
17	Direct investments			
18	Other private long-term claims	100 (iii)		−100
19	Private short-term claims	50 (ia), 5 (v)	50 (ib)	−5
20	**Changes in foreign claims on residents**			**195**
21	Foreign official claims		20 (iv)	20
22	Foreign private long-term claims			
23	Foreign private short-term claims	20 (iv), 50 (ib)	45 (ii), 100 (iii), 100 (vi)	175
24	**Other**	100 (vi)		**−100**
	Total	~~270~~ 370	~~270~~ 370	**0**
	Current account: (1) + (5) + (9)			**10**
	Capital account: (24)			**−100**
	Financial account: (10) + (20)			**90**

4.3.2. Commercial Imports: Transaction (ii)

A German utility company imports gas from Russia valued at EUR45 million (ii), and agrees to pay the Russian company within three months. The imported gas generates a debit on line 6. The obligation to pay is recorded as a credit to foreign private short-term claims on line 23.

4.3.3. Loans to Borrowers Abroad: Transaction (iii)

A German commercial bank purchases EUR100 million in intermediate-term bonds issued by a Ukrainian steel company. The bonds are denominated in euros, so payment is made in euros (i.e., by transferring EUR demand deposits). A debit entry on line 18 records the increase in German holdings of Ukrainian bonds, and a credit entry on line 23 records the increase in demand deposits held by Ukrainians in German banks.

4.3.4. Purchases of Home-Country Currency by Foreign Central Banks: Transaction (iv)

Private foreigners may not wish to retain euro balances acquired in earlier transactions. Those who are holding foreign currency, in our example euro claims, typically do so for purposes of financing purchases from Germany (or other euro area member countries). Assume, for instance, that Swiss residents attempt to sell EUR20 million in exchange for their native currency, the Swiss franc (CHF), but there is a lack of demand for EUR funds in Switzerland. In such circumstances, the CHF would appreciate against the EUR. To prevent an undesired CHF appreciation, the Swiss National Bank (SNB) might sell CHF in exchange for EUR balances.

Suppose that the Swiss National Bank purchased EUR20 million, typically in the form of a EUR demand deposit held with a German bank, from local commercial banks in Switzerland. The German balance of payments would register an increase of EUR20 million in German liabilities held by foreign monetary authorities, the Swiss National Bank (line 21), and an equivalent decline in short-term liabilities held by private foreigners (i.e., Swiss private investors, line 23). It may be noteworthy that when the SNB purchases EUR funds from Swiss commercial banks, it also credits them the CHF equivalent of EUR20 million. The SNB's liabilities to Swiss commercial banks arising from this transaction are in fact reserve deposits that Swiss banks can use when they expand their lending business and create new deposits. Currency interventions by central banks, therefore, can contribute to an increase in a country's overall money supply, all else remaining unchanged.

4.3.5. Receipts of Income from Foreign Investments: Transaction (v)

Each year, residents of Germany receive billions of EUR in interest and dividends from capital invested in foreign securities and other financial claims. German residents receive these payments in return for allowing foreigners to use German capital that otherwise could be put to work in Germany. Foreign residents, in turn, receive similar returns for their investments in Germany. Assume that a German firm has a long-term capital investment in a profitable subsidiary abroad, and that the subsidiary transfers to its German parent EUR5 million in dividends in the form of funds held in a foreign bank. The German firm then has a new (or increased) demand deposit in a foreign bank as compensation for allowing its capital to be used by its subsidiary. A debit entry on line 19 shows German private short-term claims on foreigners have increased by EUR5 million, and a credit entry on line 4 reflects the fact that German residents have given up an asset (the services of capital covered over the period) valued at EUR5 million.

4.3.6. Purchase of Nonfinancial Assets: Transaction (vi)

In a move to safeguard its long-term supply of uranium, a German utility company purchases from the government of Kazakhstan the rights to exploit a uranium mine. It agrees to pay within three months. The respective entries are on lines 23 and 24. Because a nonfinancial, nonproduced asset is involved in this transaction, it is recorded in Germany's capital account.

Note that the sum of all BOP entries in Exhibit 8-16 is 0. Transactions (i)–(iv) produce a current account surplus of EUR10 million, a capital account deficit of EUR100 million, and a financial account surplus of EUR90 million.

Although it is important to understand the detailed structure of official balance of payments accounts as described in the preceding paragraphs, this example is not necessarily how investment professionals think about the balance of payments day to day. Practitioners often think of the current account as roughly synonymous with the trade balance (merchandise trade + services) and lump all the financing flows (financial account + capital account) into one category that is usually referred to simply as the capital account. They then think of the capital account as consisting of two types of flows—portfolio investment flows and foreign direct investment (FDI). The former are shorter-term investments in foreign assets (stocks, bonds, etc.), whereas the latter are long-term investments in production capacity abroad. Although not completely accurate, this way of thinking about the balance of payments focuses attention on the components—trade, portfolio flows, and FDI—that are most sensitive to, and most likely to affect, market conditions, prices of goods and services, asset prices, and exchange rates. In addition, this perspective fits well with the role that the balance of payments plays in the macroeconomy.

4.4. National Economic Accounts and the Balance of Payments

In a closed economy, all output Y is consumed or invested by the private sector—domestic households and businesses—or purchased by the government. Letting Y denote GDP, C private consumption, I investment, and G government purchases of goods and services, the national income identity for a closed economy is given by:

$$Y = C + I + G \tag{8-1}$$

Once foreign trade is introduced, however, some output is purchased by foreigners (exports) whereas some domestic spending is used for purchases of foreign goods and services (imports). The national income identity for an open economy is thus:

$$Y = C + I + G + X - M \tag{8-2}$$

where X denotes exports and M denotes imports.

For most countries, exports rarely equal imports. Net exports or the difference between exports and imports ($X - M$) is the equivalent of the current account balance from a BOP perspective.[25] When a country's imports exceed its exports, the current account is in deficit.

[25]Strictly speaking, net exports as defined here are the trade balance rather than the current account balance because they exclude income receipts and unilateral transfers. This distinction arises because we have defined income Y as GDP rather than GNP (see section 2.1). Because the trade balance is usually the dominant component of the current account, the terms *trade balance* and *current account* are often used interchangeably. We will do so here unless the distinction is important to the discussion.

When a country's exports exceed its imports, the current account is in surplus. As the right side of Equation 8-2 shows, a current account surplus or deficit can affect GDP (and also employment). The balance of the current account is also important because it measures the size and direction of international borrowing.

In order for the balance of payments to balance, a deficit or surplus in the current account must be offset by an opposite balance in the sum of the capital and financial accounts. This requirement means that a country with a current account deficit has to increase its net foreign debts by the amount of the current account deficit. For example, the United States has run current account deficits for many years while accumulating net foreign liabilities: The current account deficit was financed by net capital imports (i.e., direct investments by foreigners), loans by foreign banks, and the sale of U.S. equities and fixed-income securities to foreign investors. By the same token, an economy with a current account surplus is earning more for its exports than it spends for its imports. Japan, Germany, and China are traditional current account surplus countries accumulating substantial net foreign claims, especially against the United States. An economy with a current account surplus finances the current account deficit of its trading partners by lending to them—that is, granting bank loans and investing in financial and real assets. As a result, the foreign wealth of a surplus country rises because foreigners pay for imports by issuing liabilities that they will eventually have to redeem.

By rearranging Equation 8-2, we can define the current account balance from the perspective of the national income accounts as:

$$CA = X - M = Y - (C + I + G) \tag{8-3}$$

Only by borrowing money from foreigners can a country have a current account deficit and consume more output than it produces. If it consumes less output than it produces, it has a current account surplus and can (indeed must) lend the surplus to foreigners. International capital flows essentially reflect an *intertemporal trade.* An economy with a current account deficit is effectively importing present consumption and exporting future consumption.

Let us now turn to the relationship between output Y and disposable income Y^d. We have to recognize that part of income is spent on taxes T, and that the private sector receives net transfers R in addition to (national) income. Disposable income Y^d is thus equal to income plus transfers minus taxes:

$$Y^d = Y + R - T \tag{8-4}$$

Disposable income, in turn, is allocated to consumption and saving, so we can write:

$$Y^d = C + S_p \tag{8-5}$$

where S_p denotes private sector saving. Combining Equations 8-4 and 8-5 allows us to write consumption as income plus transfers minus taxes and saving:

$$C = Y^d - S_p = Y + R - T - S_p \tag{8-6}$$

We can now use the right side of Equation 8-6 to substitute for C in Equation 8-3. With some rearrangement we obtain:

$$CA = S_p - I + (T - G - R) \tag{8-7}$$

Because $(T - G - R)$ is taxes minus government spending and transfers, it is the government surplus or, put differently, government savings S_g. Equation 8-7 can therefore be restated as:

$$S_p + S_g = I + CA \tag{8-8}$$

Equation 8-8 highlights an essential difference between open and closed economies: An open economy can use its saving for domestic investment or for foreign investment (i.e., by exporting its savings and acquiring foreign assets), whereas in a closed economy savings can be used only for domestic investment. Put another way, an open economy with promising investment opportunities is not constrained by its domestic savings rate in order to exploit these opportunities. As Equation 8-8 shows, it can raise investment by increasing foreign borrowing (a reduction in *CA*) without increasing domestic savings. For example, if India decides to build a network of high-speed trains, it can import all the required materials it needs from France and then borrow the funds, perhaps also from France, to pay for the materials. This transaction increases India's domestic investment because the imported materials contribute to the expansion in the country's capital stock. All else being equal, this transaction will also produce a current account deficit for India by an amount equal to the increase in investment. India's savings does not have to increase, even though investment increases. This example can be interpreted as an intertemporal trade, in which India imports present consumption (when it borrows to fund current expenditure) and exports future consumption (when it repays the loan).

Rearranging Equation 8-8, we can write:

$$S_p = I + CA - S_g \tag{8-9}$$

Equation 8-9 states that an economy's private savings can be used in three ways: (1) investment in domestic capital (*I*), (2) purchases of assets from foreigners (*CA*), and (3) net purchases (or redemptions) of government debt $(-S_g)$.

Finally, we can rearrange Equation 8-8 again to illustrate the macroeconomic sources of a current account imbalance:

$$CA = S_p + S_g - I \tag{8-10}$$

A current account deficit tends to result from low private savings, high private investment, a government deficit ($S_g < 0$), or a combination of the three. Alternatively, a current account surplus reflects high private savings, low private investment, or a government surplus.

As outlined earlier, trade deficits can result from a lack of private or government savings or booming investments. If trade deficits primarily reflect high private or government consumption (i.e., scarce savings $= S_p + S_g$), the deficit country's capacity to repay its liabilities from future production remains unchanged. If a trade deficit primarily reflects strong investments (*I*), however, the deficit country can increase its productive resources and its ability to repay its liabilities.

We can also see from Equation 8-3 that a current account deficit tends to reflect a strong domestic economy (elevated consumer, government, and investment spending), which is usually accompanied by elevated domestic credit demand and high interest rates. In such an environment, widening interest rate differentials vis-à-vis other countries can lead to growing net capital imports and produce an appreciating currency. In the long run, however, a persistent current account deficit leads to a permanent increase in the claims held by other countries against the deficit country. As a result, foreign investors may require rising risk premiums for such claims, a process that appears to lead to a depreciating currency.

EXAMPLE 8-11 The United Kingdom Budget

A financial newspaper had the following item:

> The UK's budget deficit is the highest in the G-20; in Europe, only Ireland borrows more. These are the stark facts facing Chancellor of the Exchequer George Osborne as he plans his first Budget tomorrow. He intends to tackle the problem even if that involves severe spending cuts and large tax increases.
>
> *Source: Financial Times*, 21 June 2010.

1. What are the likely consequences for the UK current account balance from the planned fiscal policy moves mentioned in the article?
2. Describe the impact spending cuts and tax increases are likely to have on UK imports.

Solution to 1: The combination of spending cuts and tax increases will, all else remaining the same, lead to an improvement in the UK current account position.

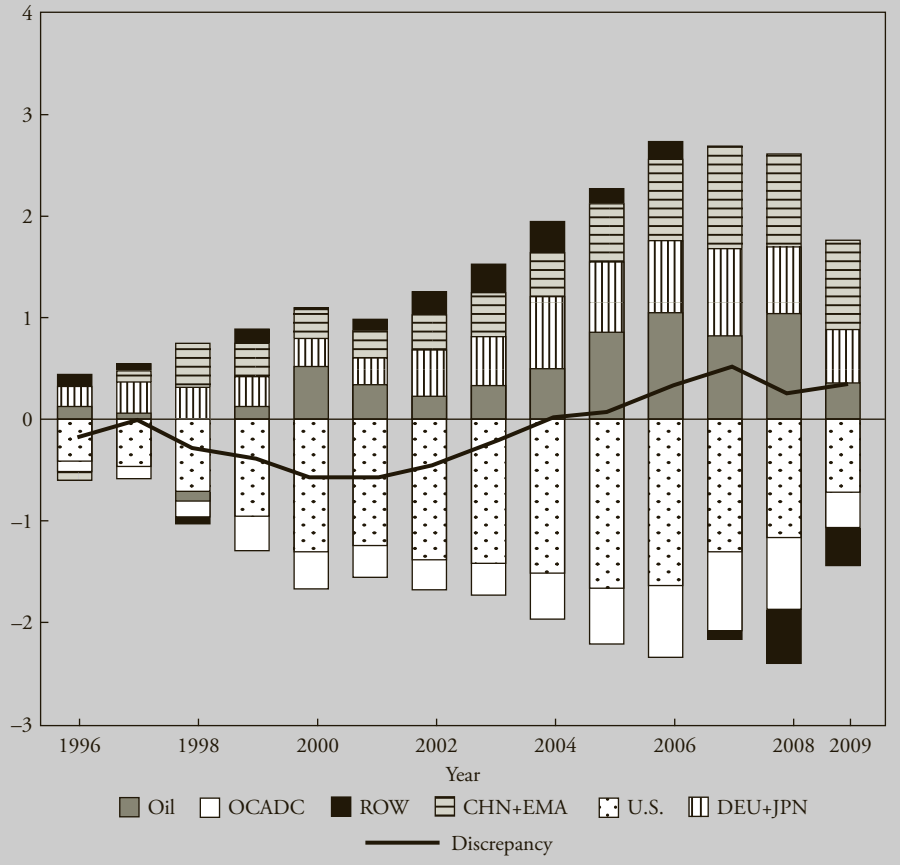
Solution to 2: UK imports are likely to be reduced by tax increases and spending cuts because government demand for foreign goods will fall and growth in private household income, which finances private imports, will be restricted as more household income goes to taxes.

Global Current Account Imbalances since 1996

As a result of growing financial integration and trade liberalization, the world economy has entered a period of rapid growth in cross-border trade since the late 1980s. In sync with surging international trade, current account imbalances widened substantially in the 1990s and the first decade of the new millennium. Exhibit 8-17 shows current account balances for 1996 to 2009 for four specific country groups—the United States, oil exporters, Germany and Japan (DEU+JPN), China and emerging Asia (CHN+EMA)—and three broad categories: the rest of the world (ROW); other

current account deficit countries (OCADC), which includes Central and Eastern European countries, Australia, New Zealand, and a wide range of smaller developed and emerging countries; and oil exporting nations. The United States ran a current account deficit in every year, and in every year its deficit represented most of the aggregate value of such deficits worldwide. Only recently, in the wake of the 2007–2009 recession, has the U.S. deficit declined both in absolute terms and relative to the global aggregate of current account deficits. In the first half of the 1990s, Germany and Japan were the traditional current account surplus countries, providing net exports of goods and services to and accumulating net claims against the United States. In the late 1990s, China and other emerging Asian nations and oil exporting nations became new and even more important surplus countries.

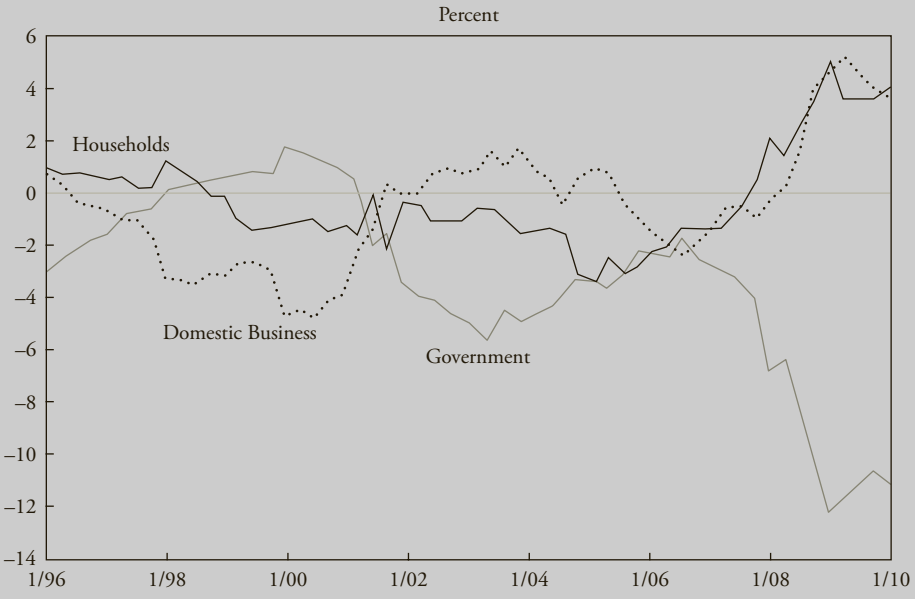
EXHIBIT 8-17 Global Imbalances (Current Account Balance in Percentage of World GDP)



Note: CHN+EMA includes China, Hong Kong SAR, Indonesia, Korea, Malaysia, Philippines, Singapore, Taiwan, and Thailand.
Source: International Monetary Fund (2010c, ch. 4).

As illustrated in Equation 8-10, current account deficits or surpluses reflect imbalances between national savings (including government savings) and investments. Current account deficits are often related to expansionary fiscal policy and government deficits. In the 1980s, for instance, the growing deficit in the U.S. current account was widely seen as the consequence of tax cuts and rising defense spending adopted by the Reagan administration. Since the mid-1990s, however, the current account imbalances depicted in Exhibit 8-17 appear to reflect other, more complex factors. Exhibit 8-18 illustrates U.S. net savings (*S*–*I*) for private domestic businesses, households, and the government (i.e., federal, state, and local) from 1996 to the second quarter of 2010. The exhibit indicates that business-sector net savings and government net savings as a percentage of GDP have been near mirror images since 1996. During the technology bubble, businesses invested heavily and ran progressively larger savings deficits while the government moved to a surplus. After the bubble burst, the pattern reversed with businesses moving to net positive savings and the government fiscal balance deteriorating sharply. Meanwhile, the household sector gradually reduced its savings rate. By 2006, each of the three sectors was dis-saving roughly 2 percent of U.S. GDP. From that point, the public and private sectors diverged sharply. In the wake of the global financial crisis, households and businesses cut spending and increased savings sharply while the government deficit exploded to more than 12 percent of GDP.

EXHIBIT 8-18 United States: Sectorial Saving–Investment Balance since 1996 (Net Savings as Percentage of GDP)



Source: Federal Reserve Board, flow of funds data.

5. TRADE ORGANIZATIONS

During the Great Depression in the 1930s, countries attempted to support their failing economies by sharply raising barriers to foreign trade, devaluing their currencies to compete against each other for export markets, and restricting their citizens' freedom to hold foreign exchange. These attempts proved to be self-defeating. World trade declined dramatically, and employment and living standards fell sharply in many countries. By the 1940s, it had become a widespread conviction that the world economy was in need of organizations that would help promote international economic cooperation. In July 1944, during the United Nations Monetary and Financial Conference in Bretton Woods, New Hampshire, representatives of 45 governments agreed on a framework for international economic cooperation. Two crucial multinational organizations emanated from this conference—the World Bank, which was founded during the conference, and the International Monetary Fund (IMF), which came into formal existence in December 1945. Although the IMF was founded with the goal to stabilize exchange rates and assist the reconstruction of the world's international payment system, the World Bank was created to facilitate postwar reconstruction and development.

A third institution, the International Trade Organization (ITO), was to be created to handle the trade side of international economic cooperation, joining the other two Bretton Woods institutions. The draft ITO charter was ambitious, extending beyond world trade regulations to include rules on employment, commodity agreements, restrictive business practices, international investment, and services. The objective was to create the ITO at a United Nations Conference on Trade and Employment in Havana, Cuba, in 1947. Meanwhile, 15 countries had begun negotiations in December 1945 to reduce and regulate customs tariffs. With World War II only barely ended, they wanted to give an early boost to trade liberalization and begin to correct the legacy of protectionist measures that had remained in place since the early 1930s. The group had expanded to 23 nations by the time the deal was signed on 30 October 1947 and the General Agreement on Tariffs and Trade (GATT) was born. The Havana conference began on 21 November 1947, less than a month after GATT was signed. The ITO charter was finally approved in Havana in March 1948, but ratification in some national legislatures proved impossible. The most serious opposition was in the U.S. Congress, even though the U.S. government had been one of the driving forces. In 1950, the United States government announced that it would not seek congressional ratification of the Havana charter, and the ITO was effectively dead. As a consequence, the GATT became the only multilateral instrument governing international trade from 1948 until the World Trade Organization (WTO) was officially established in 1995.

5.1. International Monetary Fund

As we saw earlier, current account deficits reflect a shortage of net savings in an economy and can be addressed by policies designed to rein in domestic demand. This approach could, however, have adverse consequences for domestic employment. The IMF stands ready to lend foreign currencies to member countries to assist them during periods of significant external deficits. A pool of gold and currencies contributed by members provides the IMF with the resources required for these lending operations. The funds are lent only under strict conditions, and borrowing countries' macroeconomic policies are continually monitored. The IMF's main mandate is to ensure the stability of the international monetary system, the system

of exchange rates and international payments that enables countries to buy goods and services from each other. More specifically, the IMF:

- Provides a forum for cooperation on international monetary problems.
- Facilitates the growth of international trade and promotes employment, economic growth, and poverty reduction.
- Supports exchange rate stability and an open system of international payments.
- Lends foreign exchange to members when needed, on a temporary basis and under adequate safeguards, to help them address balance of payments problems.

The global financial crisis of 2007 to 2009 demonstrated that domestic and international financial stability cannot be taken for granted, even in the world's most developed countries. In light of these events, the IMF has redefined and deepened its operations by:[26]

- *Enhancing its lending facilities.* The IMF has upgraded its lending facilities to better serve its members. As part of a wide-ranging reform of its lending practices, it has also redefined the way it engages with countries on issues related to structural reform of their economies. In this context, it has doubled member countries' access to fund resources and streamlined its lending approach to reduce the stigma of borrowing for countries in need of financial help.
- *Improving the monitoring of global, regional, and country economies.* The IMF has taken several steps to improve economic and financial surveillance, which is its framework for providing advice to member countries on macroeconomic policies and warning member countries of risks and vulnerabilities in their economies.
- *Helping resolve global economic imbalances.* The IMF's analysis of global economic developments provides finance ministers and central bank governors with a common framework for discussing the global economy.
- *Analyzing capital market developments.* The IMF is devoting more resources to the analysis of global financial markets and their links with macroeconomic policy. It also offers training to country officials on how to manage their financial systems, monetary and exchange regimes, and capital markets.
- *Assessing financial sector vulnerabilities.* Resilient, well-regulated financial systems are essential for macroeconomic stability in a world of ever-growing capital flows. The IMF and the World Bank jointly run an assessment program aimed at alerting countries to vulnerabilities and risks in their financial sectors.

From an investment perspective, the IMF helps to keep country-specific market risk and global systemic risk under control. The Greek sovereign debt crisis, which threatened to destabilize the entire European banking system, is a recent example. In early 2010, the Greek sovereign debt rating was downgraded to below investment grade by leading rating agencies as a result of serious concerns about the sustainability of Greece's public-sector debt load. Yields on Greek government bonds rose substantially following the downgrading, and the country's ability to refinance its national debt was seriously questioned in international capital markets. Bonds issued by some other European governments fell, and equity markets worldwide declined in response to spreading concerns of a Greek debt default. The downgrading of Greek sovereign debt was the ultimate consequence of persistent and growing budget deficits

[26]Visit www.imf.org/ for more information.

the Greek government had run before and after the country had joined the European Monetary Union (EMU) in 2001. Most of the budget shortfalls reflected elevated outlays for public-sector jobs, pensions, and other social benefits as well as persistent tax evasion. Reports that the Greek government had consistently and deliberately misreported the country's official economic and budget statistics contributed to further erosion of confidence in Greek government bonds in international financial markets. Facing default, the Greek government requested that a joint European Union/IMF bailout package be activated, and a loan agreement was reached among Greece, the other EMU member countries, and the IMF. The deal consisted of an immediate EUR45 billion in loans to be provided in 2010, with more funds to be available later. A total of EUR110 billion was agreed to, depending on strict economic policy conditions that included cuts in wages and benefits, an increase in the retirement age for public-sector employees, limits on public pensions, increases in direct and indirect taxes, and a substantial reduction in state-owned companies. By providing conditional emergency lending facilities to the Greek government and designing a joint program with the European Union on how to achieve fiscal consolidation, the IMF prevented a contagious wave of sovereign debt crises in global capital markets.

Another example of IMF activities is seen in the East Asian financial crisis in the late 1990s. It began in July 1997, when Thailand was forced to abandon its currency's peg with the U.S. dollar. Currency devaluation subsequently hit other East Asian countries that had similar balance of payment problems, such as South Korea, Malaysia, the Philippines, and Indonesia. They had run persistent and increasing current account deficits, financed mainly with short-term capital imports—in particular, domestic banks borrowing in international financial markets. External financing was popular because of the combination of lower foreign (especially U.S.) interest rates and fixed exchange rates. Easy money obtained from abroad led to imprudent investment, which contributed to overcapacities in several industries and inflated prices on real estate and stock markets. The IMF came to the rescue of the affected countries with considerable loans, accompanied by policies designed to control domestic demand, which included fiscal austerity and tightened monetary reins.

5.2. World Bank Group

The World Bank's main objective is to help developing countries fight poverty and enhance environmentally sound economic growth. For developing countries to grow and attract business, they have to:

- Strengthen their governments and educate their government officials.
- Implement legal and judicial systems that encourage business.
- Protect individual and property rights and honor contracts.
- Develop financial systems robust enough to support endeavors ranging from microcredit to financing larger corporate ventures.
- Combat corruption.

Given these targets, the World Bank provides funds for a wide range of projects in developing countries worldwide along with financial and technical expertise aimed at helping those countries reduce poverty.

The World Bank's two closely affiliated entities—the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA)—provide low- or no-interest loans and grants to countries that have unfavorable or no access to

international credit markets. Unlike private financial institutions, neither the IBRD nor the IDA operates for profit. The IBRD is market-based, and uses its high credit rating to pass the low interest it pays for funds on to its borrowers—developing countries. It pays for its own operating costs because it does not look to outside sources to furnish funds for overhead.

IBRD lending to developing countries is primarily financed by selling AAA-rated bonds in the world's financial markets. Although the IBRD earns a small margin on this lending, the greater proportion of its income comes from lending out its own capital. This capital consists of reserves built up over the years and money paid in from the Bank's 185 member country shareholders. The IBRD's income also pays for World Bank operating expenses and has contributed to the IDA and debt relief. The IDA is the world's largest source of interest-free loans and grant assistance to the poorest countries. Its funds are replenished every three years by 40 donor countries. Additional funds are regenerated through repayments of loan principal on 35- to 40-year, no-interest loans, which are then available for relending. At the end of September 2010, the IBRD had net loans outstanding of USD125.5 billion, while its borrowings amounted to USD132 billion.

Besides acting as a financier, the World Bank also provides analysis, advice, and information to its member countries to enable them to achieve the lasting economic and social improvements their people need. Another of the World Bank's core functions is to increase the capabilities of its partners, people in developing countries, and its own staff. It has set up links to a wide range of knowledge-sharing networks to address the vast need for information and dialogue about development.

From an investment perspective, the World Bank helps to create the basic economic infrastructure that is essential for the creation of domestic financial markets and a well-functioning financial industry in developing countries. Moreover, the IBRD is one of the most important supranational borrowers in the international capital markets. Because of its strong capital position and its very conservative financial, liquidity, and lending policies, it enjoys the top investment-grade rating from the leading agencies and investors have confidence in its ability to withstand adverse events. As a result, IBRD bonds denominated in various major currencies are widely held by institutional and private investors.

5.3. World Trade Organization

The WTO provides the legal and institutional foundation of the multinational trading system. It is the only international organization that regulates cross-border trade relationships among nations on a global scale. It was founded on 1 January 1995, replacing the General Agreement on Tariffs and Trade (GATT) that had come into existence in 1947. The GATT was the only multilateral body governing international trade from 1947 to 1995. It operated for almost half a century as a quasi-institutionalized, provisional system of multilateral treaties. Several rounds of negotiations took place under the GATT, of which the Tokyo round and the Uruguay round may have been the most far-reaching. The Tokyo round was the first major effort to address a wide range of nontariff trade barriers, whereas the Uruguay round focused on the extension of the world trading system into several new areas, particularly trade in services and intellectual property, but also to reform trade in agricultural products and textiles. The GATT still exists in an updated 1994 version and is the WTO's principal treaty for trade in goods. The GATT and the General Agreement on Trade in Services (GATS) are the major agreements within the WTO's body of treaties that encompasses a total of about 60 agreements, annexes, decisions, and understandings.

In November 2001, the most recent and still ongoing round of negotiations was launched by the WTO in Doha, Qatar. The Doha round was an ambitious effort to enhance globalization by slashing barriers and subsidies in agriculture and addressing a wide range of cross-border services. So far, under GATS, which came into force in January 1995, banks, insurance companies, telecommunication firms, tour operators, hotel chains, and transport companies that want to do business abroad can enjoy the same principles of free and fair trade that had previously applied only to international trade in goods. No agreement has been reached in Doha so far, however, despite intense negotiations at several ministerial conferences and at other sessions. The start of the Doha round nevertheless marks one of the most crucial events in global trade over the past decade: China's accession to the WTO in December 2001.

The WTO's most important functions are the implementation, administration, and operation of individual agreements; acting as a platform for negotiations; and settling disputes. Moreover, the WTO has the mandate to review and propagate its members' trade policies and ensure the coherence and transparency of trade policies through surveillance in a global policy setting. The WTO also provides technical cooperation and training to developing, least developed, and low-income countries to assist with their adjustment to WTO rules. In addition, the WTO is a major source of economic research and analysis, producing ongoing assessments of global trade in its publications and research reports on special topics. Finally, the WTO is in close cooperation with the other two Bretton Woods institutions, the IMF and the World Bank.

From an investment perspective, it would be hard to conceive of today's global multinational corporations without the major institutional and regulatory base provided by the WTO's framework of global trade rules. Modern financial markets would look different without the large, multinational companies whose stocks and bonds have become key elements in investment portfolios. In the equity universe, for instance, investment considerations focusing on global sectors rather than national markets would make little sense without a critical mass of multinational firms competing with each other in a globally defined business environment.

EXAMPLE 8-12 Function and Objective of International Organizations

On 10 May 2010, the Greek government officially applied for emergency lending facilities extended by the International Monetary Fund. It sent the following letter to Dominique Strauss-Kahn, the IMF's managing director:

> **Request for Stand-By Arrangement**
> This paper was prepared based on the information available at the time it was completed on Monday, May 10, 2010. The views expressed in this document are those of the staff team and do not necessarily reflect the views of the government of Greece or the Executive Board of the IMF. The policy of publication of staff reports and other documents by the IMF allows for the deletion of market-sensitive information.

Mr. Dominique Strauss-Kahn
Managing Director
International Monetary Fund
Washington DC

Athens, May 3, 2010

Dear Mr. Strauss-Kahn:

The attached Memorandum of Economic and Financial Policies (MEFP)* outlines the economic and financial policies that the Greek government and the Bank of Greece, respectively, will implement during the remainder of 2010 and in the period 2011–2013 to strengthen market confidence and Greece's fiscal and financial position during a difficult transition period toward a more open and competitive economy. The government is fully committed to the policies stipulated in this document and its attachments, to frame tight budgets in the coming years with the aim to reduce the fiscal deficit to below 3 percent in 2014 and achieve a downward trajectory in the public debt–GDP ratio beginning in 2013, to safeguard the stability of the Greek financial system, and to implement structural reforms to boost competitiveness and the economy's capacity to produce, save, and export. (. . .) The government is strongly determined to lower the fiscal deficit, (. . .) by achieving higher and more equitable tax collections, and constraining spending in the government wage bill and entitlement outlays, among other items. In view of these efforts and to signal the commitment to effective macroeconomic policies, the Greek government requests that the Fund supports this multi-year program under a Stand-By Arrangement (SBA) for a period of 36 months in an amount equivalent to SDR26.4 billion.† (. . .) A parallel request for financial assistance to euro area countries for a total amount of €80 billion has been sent. The implementation of the program will be monitored through quantitative performance criteria and structural benchmarks as described in the attached MEFP and Technical Memorandum of Understanding (TMU). There will be twelve quarterly reviews of the program supported under the SBA by the Fund, (. . .) to begin with the first review that is expected to be completed in the course of the third calendar quarter of 2010, and then every quarter thereafter until the last quarterly review envisaged to be completed during the second calendar quarter of 2013, to assess progress in implementing the program and reach understandings on any additional measures that may be needed to achieve its objectives. (. . .) The Greek authorities believe that the policies set forth in the attached memorandum are adequate to achieve the objectives of the economic program, and stand ready to take any further measures that may become appropriate for this purpose. The authorities will consult with the Fund in accordance with its policies on such consultations, (. . .) and in advance of revisions to the policies contained in the MEFP. All information requested by the Fund (. . .) to assess implementation of the program will be provided.

(. . .)

Sincerely,

George Papaconstantinou
Minister of Finance

George Provopoulos
Governor of the Bank of Greece

1. What is the objective of the IMF's emergency lending facilities?
2. What are the macroeconomic policy conditions under which the IMF provides emergency lending to Greece?
3. What is the amount Greece requests from the IMF as emergency funds?

Solution to 1: The program seeks to safeguard the stability of the Greek financial system and to implement structural reforms to boost competitiveness and the economy's capacity to produce, save, and export.

Solution to 2: The Greek government has to reduce the country's fiscal deficit by achieving higher and more equitable tax collections as well as constrain spending in the government wage bill and entitlement outlays.

Solution to 3: Greece applied for a standby arrangement in an amount equivalent to SDR26.4 billion (approximately USD39.5 billion, based on the 10 May 2010 exchange rate).

*The detailed memorandum is available from www.imf.org/external/pubs/ft/scr/2010/cr10111.pdf.

†An SDR (special drawing right) is a basket of four leading currencies: Japanese yen (JPY), U.S. dollar (USD), British pound (GBP), and euro (EUR). It consists of 18.4 yen, 0.6320 USD, 0.0903 GBP, and 0.41 EUR. One SDR was worth 1.4975 USD or 1.1547 EUR on 10 May 2010.

6. SUMMARY

This chapter provides a framework for analyzing a country's trade and capital flows and their economic implications. It examines basic models that explain trade based on comparative advantage and provides a basis for understanding how international trade can affect the rate and composition of economic growth as well as the attractiveness of investment in various sectors.

- The benefits of trade include:
  - Gains from exchange and specialization.
  - Gains from economies of scale as companies add new markets for their products.
  - Greater variety of products available to households and firms.
  - Increased competition and more efficient allocation of resources.
- A country has an absolute advantage in producing a good (or service) if it is able to produce that good at a lower absolute cost or use fewer resources in its production than its trading partner. A country has a comparative advantage in producing a good if its *opportunity cost* of producing that good is less than that of its trading partner.
- Even if a country does not have an absolute advantage in the production of any good, it can gain from trade by producing and exporting the good(s) in which it has a comparative advantage and importing good(s) in which it has a comparative disadvantage.
- In the Ricardian model of trade, comparative advantage and the pattern of trade are determined by differences in technology between countries. In the Heckscher–Ohlin model

of trade, comparative advantage and the pattern of trade are determined by differences in factor endowments between countries. In reality, technology and factor endowments are complementary, not mutually exclusive, determinants of trade patterns.

- Trade barriers prevent the free flow of goods and services among countries. Governments impose trade barriers for various reasons, including to promote specific developmental objectives, to counteract certain imperfections in the functioning of markets, or to respond to problems facing their economies.
- For purposes of international trade policy and analysis, a small country is defined as one that cannot affect the world price of traded goods. A large country's production and/or consumption decisions do alter the relative prices of trade goods.
- In a small country, trade barriers generate a net welfare loss arising from distortion of production and consumption decisions and the associated inefficient allocation of resources.
- Trade barriers can generate a net welfare gain in a large country if the gain from improving its terms of trade (higher export prices and lower import prices) more than offsets the loss from the distortion of resource allocations. However, the large country can gain only if it imposes an even larger welfare loss on its trading partner(s).
- An import tariff and an import quota have the same effect on price, production, and trade. With a quota, however, some or all of the revenue that would be raised by the equivalent tariff is instead captured by foreign producers (or the foreign government) as quota rents. Thus, the welfare loss suffered by the importing country is generally greater with a quota.
- A voluntary export restraint is imposed by the exporting country. It has the same impact on the importing country as an import quota from which foreigners capture all of the quota rents.
- An export subsidy encourages firms to export their product rather than sell it in the domestic market. The distortion of production, consumption, and trade decisions generates a welfare loss. The welfare loss is greater for a large country because increased production, and export, of the subsidized product reduces its global price—that is, worsens the country's terms of trade.
- Capital restrictions are defined as controls placed on foreigners' ability to own domestic assets and/or domestic residents' ability to own foreign assets. In contrast to trade restrictions, which limit the openness of goods markets, capital restrictions limit the openness of financial markets.
- A regional trading bloc is a group of countries that have signed an agreement to reduce and progressively eliminate barriers to trade and movement of factors of production among the members of the bloc.
  - The bloc may or may not have common trade barriers against those countries that are not members of the bloc. In a free trade area all barriers to the flow of goods and services among members are eliminated, but each country maintains its own polices against nonmembers.
  - A customs union extends the FTA by not only allowing free movement of goods and services among members but also creating a common trade policy against nonmembers.
  - A common market incorporates all aspects of a customs union and extends it by allowing free movement of factors of production among members.
  - An economic union incorporates all aspects of a common market and requires common economic institutions and coordination of economic policies among members.
  - Members of a monetary union adopt a common currency.
- From an investment perspective, it is important to understand the complex and dynamic nature of trading relationships because they can help identify potential profitable investment opportunities as well as provide some advance warning signals regarding when to disinvest in a market or industry.

- The major components of the balance of payments are:
  - The current account balance, which largely reflects trade in goods and services.
  - The capital account balance, which mainly consists of capital transfers and net sales of nonproduced, nonfinancial assets.
  - The financial account, which measures net capital flows based on sales and purchases of domestic and foreign financial assets.
- Decisions by consumers, firms, and governments influence the balance of payments.
  - Low private savings and/or high investment tend to produce a current account deficit that must be financed by net capital imports; high private savings and/or low investment, however, produce a current account surplus, balanced by net capital exports.
  - All else being the same, a government deficit produces a current account deficit and a government surplus leads to a current account surplus.
  - All else being the same, a sustained current account deficit contributes to a rise in the risk premium for financial assets of the deficit country. Current account surplus countries tend to enjoy lower risk premiums than current account deficit countries.
- Created after World War II, the International Monetary Fund, the World Bank, and the World Trade Organization are the three major international organizations that provide necessary stability to the international monetary system and facilitate international trade and development.
  - The IMF's mission is to ensure the stability of the international monetary system, the system of exchange rates and international payments that enables countries to buy goods and services from each other. The IMF helps to keep country-specific market risk and global systemic risk under control.
  - The World Bank helps to create the basic economic infrastructure essential for creation and maintenance of domestic financial markets and a well-functioning financial industry in developing countries.
  - The World Trade Organization's mission is to foster free trade by providing a major institutional and regulatory framework of global trade rules; without this framework it would be hard to conceive of today's global multinational corporations.

PRACTICE PROBLEMS[27]

1. Which of the following statements *best* describes the benefits of international trade?
   A. Countries gain from exchange and specialization.
   B. Countries receive lower prices for their exports and pay higher prices for imports.
   C. Absolute advantage is required for a country to benefit from trade in the long term.

2. Which of the following statements *best* describes the costs of international trade?
   A. Countries without an absolute advantage in producing a good cannot benefit significantly from international trade.
   B. Resources may need to be allocated into or out of an industry, and less efficient companies may be forced to exit an industry, which in turn may lead to higher unemployment.
   C. Loss of manufacturing jobs in developed countries as a result of import competition means that developed countries benefit far less than developing countries from trade.

[27]These practice problems were developed by Drew H. Boecher, CFA (Dedham, Massachusetts, USA).

3. Suppose the cost of producing tea relative to copper is lower in Tealand than in Copperland. With trade, the copper industry in Copperland would *most likely*:
   A. expand.
   B. contract.
   C. remain stable.

4. A country has a comparative advantage in producing a good if:
   A. it is able to produce the good at a lower cost than its trading partner.
   B. its opportunity cost of producing the good is less than that of its trading partner.
   C. its opportunity cost of producing the good is more than that of its trading partner.

5. Suppose Mexico exports vegetables to Brazil and imports flashlights used for mining from Brazil. The output per worker per day in each country is as follows:

	Flashlights	Vegetables
Mexico	20	60
Brazil	40	80

   Which country has a comparative advantage in the production of vegetables, and what is the *most* relevant opportunity cost?
   A. Brazil: 2 vegetables per flashlight
   B. Mexico: 1.5 vegetables per flashlight
   C. Mexico: ⅓ flashlight per vegetable

6. Suppose three countries produce rulers and pencils with output per worker per day in each country as follows:

	Rulers	Pencils
Mexico	20	40
Brazil	30	90
China	40	160

   Which country has the greatest comparative advantage in the production of rulers?
   A. China
   B. Brazil
   C. Mexico

7. In the Ricardian trade model, comparative advantage is determined by:
   A. technology.
   B. the capital-to-labor ratio.
   C. the level of labor productivity.

8. In the Ricardian trade model, a country captures more of the gains from trade if:
   A. it produces all products while its trade partner specializes in one good.
   B. the terms of trade are closer to its autarkic prices than to its partner's autarkic prices.
   C. the terms of trade are closer to its partner's autarkic prices than to its autarkic prices.

9. Germany has much more capital per worker than Portugal. In autarky each country produces and consumes both machine tools and wine. Production of machine tools is relatively capital intensive whereas winemaking is labor intensive. According to the Heckscher–Ohlin model, when trade opens:
   A. Germany should export machine tools and Portugal should export wine.
   B. Germany should export wine and Portugal should export machine tools.
   C. Germany should produce only machine tools and Portugal should produce only wine.

10. According to the Heckscher–Ohlin model, when trade opens:
    A. the scarce factor gains relative to the abundant factor in each country.
    B. the abundant factor gains relative to the scarce factor in each country.
    C. income is redistributed between countries but not within each country.

11. Which type of trade restriction would *most likely* increase domestic government revenue?
    A. A tariff
    B. An import quota
    C. An export subsidy

12. Which of the following trade restrictions is *most likely* to result in the greatest welfare loss for the importing country?
    A. A tariff
    B. An import quota
    C. A voluntary export restraint

13. A large country can:
    A. benefit by imposing a tariff.
    B. benefit with an export subsidy.
    C. not benefit from any trade restriction.

14. If Brazil and South Africa have free trade with each other, a common trade policy against all other countries, but no free movement of factors of production between them, then Brazil and South Africa are part of a:
    A. customs union.
    B. common market.
    C. free trade area (FTA).

15. Which of the following factors *best* explains why regional trading agreements are more popular than larger multilateral trade agreements?
    A. Minimal displacement costs
    B. Trade diversions that benefit members
    C. Quicker and easier policy coordination

16. The sale of mineral rights would be captured in which of the following balance of payments components?
    A. Capital account
    B. Current account
    C. Financial account

17. Patent fees and legal services are recorded in which of the following balance of payments components?
    A. Capital account
    B. Current account
    C. Financial account

18. During the most recent quarter, a steel company in South Korea had the following transactions:
    - The company bought iron ore from Australia for AUD50 million.
    - It sold finished steel to the United States for USD65 million.
    - It borrowed AUD50 million from a bank in Sydney.
    - It received a USD10 million dividend from a U.S. subsidiary.
    - It paid KRW550 million to a South Korean shipping company.

    Which of the following would be reflected in South Korea's current account balance for the quarter?
    A. The loan
    B. The shipping
    C. The dividend

19. Which of the following *most likely* contributes to a current account deficit?
    A. High taxes
    B. Low private savings
    C. Low private investment

20. Which of the following chronic deficit conditions is *least* alarming to the deficit country's creditors?
    A. High consumption
    B. High private investment
    C. High government spending

21. Which of the following international trade organizations regulates cross-border exchange among nations on a global scale?
    A. World Bank Group (World Bank)
    B. World Trade Organization (WTO)
    C. International Monetary Fund (IMF)

22. Which of the following international trade organizations has a mission to help developing countries fight poverty and enhance environmentally sound economic growth?
    A. World Bank Group (World Bank)
    B. World Trade Organization (WTO)
    C. International Monetary Fund (IMF)

23. Which of the following organizations helps to keep global systemic risk under control by preventing contagion in scenarios such as the 2010 Greek sovereign debt crisis?
    A. World Bank Group (World Bank)
    B. World Trade Organization (WTO)
    C. International Monetary Fund (IMF)

24. Which of the following international trade bodies was the only multilateral body governing international trade from 1948 to 1995?
    A. World Trade Organization (WTO)
    B. International Trade Organization (ITO)
    C. General Agreement on Tariffs and Trade (GATT)

CHAPTER 9

CURRENCY EXCHANGE RATES

William A. Barker, CFA

Paul D. McNelis

Jerry Nickelsburg

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Define an exchange rate, and distinguish between nominal and real exchange rates and spot and forward exchange rates.
- Describe functions of and participants in the foreign exchange market.
- Calculate and interpret the percentage change in a currency relative to another currency.
- Calculate and interpret currency cross-rates.
- Convert a forward quotation expressed on a points basis or in percentage terms into an outright forward quotation.
- Explain the arbitrage relationships between spot rates, forward rates, and interest rates.
- Calculate and interpret a forward discount or premium.
- Calculate and interpret the forward rate consistent with the spot rate and the interest rate in each currency.
- Describe exchange rate regimes.
- Explain the impact of exchange rates on countries' international trade and capital flows.

1. INTRODUCTION

Measured by daily turnover, the foreign exchange (FX) market—the market in which currencies are traded against each other—is by far the world's largest market. Current estimates

put daily turnover at approximately USD4 trillion for 2010. This is about 10 to 15 times larger than daily turnover in global fixed-income markets and about 50 times larger than global turnover in equities. Moreover, volumes in FX turnover continue to grow: Some predict that daily FX turnover will reach USD10 trillion by 2020 as market participation spreads and deepens.

The FX market is also a truly global market that operates 24 hours a day, each business day. It involves market participants from every time zone connected through electronic communications networks that link players as large as multibillion-dollar investment funds and as small as individuals trading for their own accounts—all brought together in real time. International trade would be impossible without the trade in currencies that facilitates it, and so too would cross-border capital flows that connect all financial markets globally through the FX market.

These factors make foreign exchange a key market for investors and market participants to understand. The world economy is increasingly transnational in nature, with both production processes and trade flows often determined more by global factors than by domestic considerations. Likewise, investment portfolio performance increasingly reflects global determinants because pricing in financial markets responds to the array of investment opportunities available worldwide, not just locally. All of these factors funnel through, and are reflected in, the foreign exchange market. As investors shed their home bias and invest in foreign markets, the exchange rate—the price at which foreign-currency-denominated investments are valued in terms of the domestic currency—becomes an increasingly important determinant of portfolio performance.

Even investors adhering to a purely domestic portfolio mandate are increasingly affected by what happens in the foreign exchange market. Given the globalization of the world economy, most large companies depend heavily on their foreign operations (for example, by some estimates about 40 percent of S&P 500 index earnings are from outside the United States). Almost all companies are exposed to some degree of foreign competition, and the pricing for domestic assets—equities, bonds, real estate, and others—will also depend on demand from foreign investors. All of these various influences on investment performance reflect developments in the foreign exchange market.

This chapter introduces the foreign exchange market, providing the basic concepts and terminology necessary to understand exchange rates as well as some of the basics of exchange rate economics.

The chapter is divided up as follows. Section 2 describes the organization of the foreign exchange market and discusses the major players—who they are, how they conduct their business, and how they respond to exchange rate changes. Section 3 takes up the mechanics of exchange rates: definitions, quotes, and calculations. This section shows that the reader has to pay close attention to conventions used in various foreign exchange markets around the world because they can vary widely. Sometimes exchange rates are quoted in the number of domestic currency units per unit of foreign currency, and sometimes they are quoted in the opposite way. The exact notation used to represent exchange rates can vary widely as well, and occasionally the same exchange rate notation will be used by different sources to mean completely different things. The notation used here may not be the same as that encountered elsewhere. Therefore, the focus should be on understanding the underlying concepts rather than relying on rote memorization of formulas. We also show how to calculate cross-exchange rates and how to compute the forward exchange rate given either the forward points or the percentage forward premium or discount. In Section 4, we describe alternative exchange rate regimes operating throughout the world. Finally, in Section 5, we discuss how exchange rates affect a country's international trade (exports and imports) and capital flows. A summary and practice problems conclude the chapter.

2. THE FOREIGN EXCHANGE MARKET

To understand the FX market, it is necessary to become familiar with some of its basic conventions. Individual currencies are often referred to by standardized three-letter codes that the market has agreed on through the International Organization for Standardization (ISO). Exhibit 9-1 lists some of the major global currencies and their identification codes.

It is important to understand that there is a difference between referring to an *individual currency* and to an *exchange rate*. One can hold an individual currency (for example, in a EUR100 million deposit), but an exchange rate refers to the price of one currency in terms of another (for example, the exchange rate between the EUR and USD). An individual currency can be singular, but there are always two currencies involved in an exchange rate: the price of one currency relative to another. The exchange rate is the number of units of one currency (called the *price currency*) that one unit of another currency (called the *base currency*) will buy. An equivalent way of describing the exchange rate is as the cost of one unit of the base currency in terms of the price currency.

EXHIBIT 9-1 Standard Currency Codes

Three-Letter Currency Code	Currency
USD	U.S. dollar
EUR	Euro
JPY	Japanese yen
GBP	British pound
CHF	Swiss franc
CAD	Canadian dollar
AUD	Australian dollar
NZD	New Zealand dollar
ZAR	South African rand
SEK	Swedish krona
NOK	Norwegian krone
BRL	Brazilian real
SGD	Singapore dollar
MXN	Mexican peso
CNY	Chinese yuan
HKD	Hong Kong dollar
INR	Indian rupee
KRW	South Korean won
RUB	Russian ruble

This distinction between individual currencies and exchange rates is important because, as we will see in a later section, these three-letter currency codes can be used both ways. (For example, when used as an exchange rate in the professional FX market, EUR is understood to be the exchange rate between the euro and the U.S. dollar). But be aware of the context (either as a currency or as an exchange rate) in which these three-letter currency codes are being used. To avoid confusion, this chapter will identify exchange rates using the convention of "A/B," referring to the number of units of currency A that one unit of currency B will buy. For example, a USD/EUR exchange rate of 1.2875 means that one euro will buy 1.2875 U.S. dollars (i.e., one euro costs 1.2875 U.S. dollars).[1] In this case, the euro is the base currency and the U.S. dollar is the price currency. A decrease in this exchange rate would mean that the euro costs less than USD1.2875 or that a smaller amount of U.S. currency than this is needed to buy one euro. In other words, a decline in this exchange rate indicates that the USD is *appreciating* against the EUR or, equivalently, the EUR is *depreciating* against the USD.

The exchange rates just described are referred to as *nominal* exchange rates. This is to distinguish them from *real* exchange rates, which are indexes often constructed by economists and other market analysts to assess changes in the relative purchasing power of one currency compared with another. Creating these indexes requires adjusting the nominal exchange rate by using the price levels in each country of the currency pair (hence the name "real exchange rates") in order to compare the relative purchasing power between countries.

In a world of homogeneous goods and services and with no market frictions or trade barriers, the relative purchasing power across countries would tend to equalize: Why would you pay more, in real terms, domestically for a widget if you could import an identical widget from overseas at a cheaper price? This basic concept is the intuition behind a theory known as purchasing power parity (PPP), which describes the long-term equilibrium of nominal exchange rates. PPP asserts that nominal exchange rates adjust so that identical goods (or baskets of goods) will have the same price in different markets. Or, put differently, the purchasing power of different currencies is equalized for a standardized basket of goods.

In practice, the conditions required to enforce PPP are not satisfied: Goods and services are not identical across countries; countries typically have different baskets of goods and services produced and consumed; many goods and services are not traded internationally; there are trade barriers and transaction costs (e.g., shipping costs and import taxes); and capital flows are at least as important as trade flows in determining nominal exchange rates. As a result, nominal exchange rates exhibit persistent deviations from PPP. Moreover, relative purchasing power among countries displays a weak, if any, tendency toward long-term equalization. A simple example of a cross-country comparison of the purchasing power of a standardized good is the "Big Mac" index produced by the *Economist*, which shows the relative price of this standardized hamburger in different countries. The Big Mac index shows that fast-food hamburger prices can vary widely internationally and that this difference in purchasing power is typical of most goods and services. Hence, movements in real exchange rates provide meaningful information about changes in relative purchasing power among countries.

Consider the case of an individual who wants to purchase goods from a foreign country. The individual would be able to buy fewer of these goods if the nominal spot exchange rate for

[1]This convention is consistent with the meaning of "/" in mathematics, and the straightforward interpretation of "A/B" as "A per B" is helpful in understanding exchange rates as the price of one currency in terms of another. Nevertheless, other notation conventions exist. "B/A" and "B:A" are sometimes used to denote what this chapter denotes as "A/B." Careful attention to the context will usually make the convention clear.

the foreign currency appreciated or if the foreign price level increased. Conversely, the individual could buy more foreign goods if the individual's domestic income increased. (For this example, we will assume that changes in the individual's income are proportional to changes in the domestic price level.) Hence, in *real* purchasing power terms, the real exchange rate that an individual faces is an increasing function of the nominal exchange rate (quoted in terms of the number of units of domestic currency per one unit of foreign currency) and the foreign price level and a decreasing function of the domestic price level. The *higher* the real exchange rate that this individual faces, the *fewer* foreign goods, in real terms, the individual can purchase and the *lower* that individual's relative purchasing power compared with citizens in the other country.

An equivalent way of viewing the real exchange rate is that it represents the relative price levels in the domestic and foreign countries. Mathematically, we can represent the foreign price level in terms of the domestic currency as:

$$\text{Foreign price level in domestic currency} = S_{d/f} \times P_f$$

where $S_{d/f}$ is the spot exchange rate (quoted in terms of the number of units of domestic currency per one unit of foreign currency) and P_f is foreign price level quoted in terms of the foreign currency. We can define the domestic price level, in terms of the domestic currency, as P_d. Hence, the ratio between the foreign and domestic price levels is:

$$\text{Real exchange rate}_{d/f} = (S_{d/f} \times P_f)/P_d = S_{d/f} \times (P_f/P_d)$$

For example, for a British consumer wanting to buy goods made in the Eurozone, the real exchange rate (defined in GBP/EUR terms; note that the domestic currency for the United Kingdom is the price currency, not the base currency) will be an increasing function of the nominal spot exchange rate (GBP/EUR) and the Eurozone price level and a decreasing function of the UK price level. This is written as:

$$\text{Real exchange rate}_{\frac{GBP}{EUR}} = S_{\frac{GBP}{EUR}} \times \left(\frac{CPI_{EUR}}{CPI_{UK}}\right)$$

Let's examine the effect of movements in the domestic and foreign price levels, and the nominal spot exchange rate, on the real purchasing power of an individual in the United Kingdom wanting to purchase Eurozone goods. Assume that the nominal spot exchange rate (GBP/EUR) increases by 10 percent, the Eurozone price level by 5 percent, and the UK price level by 2 percent. The change in the real exchange rate is then:

$$\left(1 + \frac{\Delta S_{\frac{d}{f}}}{S_{\frac{d}{f}}}\right) \times \frac{\left(1 + \frac{\Delta P_f}{P_f}\right)}{\left(1 + \frac{\Delta P_d}{P_d}\right)} - 1 = (1 + 10\%) \times \frac{1 + 5\%}{1 + 2\%} - 1 \approx 10\% + 5\% - 2\% \approx 13\%$$

In this case, the real exchange rate for the UK-based individual has *increased* about 13 percent, meaning that it now costs *more*, in real terms, to buy Eurozone goods. Or put differently, the UK individual's real purchasing power relative to Eurozone goods has *declined* by about 13 percent. An easy way to remember this relationship is to consider the real exchange rate (stated with the domestic currency as the price currency) as representing the real price you

face in order to purchase foreign goods and services: the *higher* the price (real exchange rate), the *lower* your relative purchasing power.

The real exchange rate for a currency can be constructed for the domestic currency relative to a single foreign currency or relative to a basket of foreign currencies. In either case, these real exchange rate indexes depend on the assumptions made by the analyst creating them. Several investment banks and central banks create proprietary measures of real exchange rates. It is important to note that real exchange rates are *not* quoted or traded in global FX markets; they are only indexes created by analysts to understand the international competitiveness of an economy and the real purchasing power of a currency.

In this context, real exchange rates can be useful for understanding trends in international trade and capital flows and hence can be seen as one of the influences on nominal spot exchange rates. As an example, consider the exchange rate between the Chinese yuan and the U.S. dollar. During 2010, the nominal yuan exchange rate against the U.S. dollar (CNY/USD) declined by approximately 3 percent—meaning that the U.S. dollar depreciated against the yuan. However, the annual inflation rates in the United States and China were different during 2010—approximately 1.5 percent for the United States and 4.5 percent for China. This means that the real exchange rate (in CNY/USD terms) was depreciating more rapidly than the nominal CNY/USD exchange rate:

$$\left(1 + \%\Delta S_{\frac{CNY}{USD}}\right) \times \frac{(1 + \%\Delta P_{US})}{(1 + \%\Delta P_{China})} - 1 \approx -3\% + 1.5\% - 4.5\% \approx -6\%$$

This combination of a stronger yuan and a higher Chinese inflation rate meant that the real exchange rate faced by China was declining, thus increasing Chinese purchasing power in USD terms.

Movements in real exchange rates can have a similar effect as movements in nominal exchange rates in terms of affecting relative prices and hence trade flows. Even if the nominal spot exchange rate does not move, differences in inflation rates between countries affect their relative competitiveness.

Although real exchange rates can exert some influence on nominal exchange rate movements, they are only one of many factors; it can be difficult to disentangle all of these interrelationships in a complex and dynamic FX market. As discussed earlier, PPP is a poor guide to predicting future movements in nominal exchange rates, because these rates can deviate from PPP equilibrium—and even continue to trend away from their PPP level—for years at a time. Hence, it should not be surprising that real exchange rates, which reflect changes in relative purchasing power, have a poor track record as a predictor of future nominal exchange rate movements.

EXAMPLE 9-1 Nominal and Real Exchange Rates

An investment adviser located in Sydney, Australia, is meeting with a local client who is looking to diversify her domestic bond portfolio by adding investments in fixed-rate, long-term bonds denominated in HKD. The client frequently visits Hong Kong, and many of her annual expenses are denominated in HKD. The client, however, is

concerned about the foreign currency risks of offshore investments and whether the investment return on her HKD-denominated investments will maintain her purchasing power—both domestically (i.e., for her AUD-denominated expenses) and in terms of her foreign trips (i.e., denominated in HKD for her visits to Hong Kong). The investment adviser explains the effect of changes in nominal and real exchange rates to the client and illustrates this explanation by making the following statements:

Statement 1: "All else held equal, an increase in the nominal AUD/HKD exchange rate will lead to an increase in the AUD-denominated value of your foreign investment."

Statement 2: "All else held equal, an increase in the nominal AUD/HKD exchange rate means that your relative purchasing power for your Hong Kong trips will increase (based on paying for your trip with the income from your HKD-denominated bonds)."

Statement 3: "All else held equal, an increase in the Australian inflation rate will lead to an increase in the real exchange rate (AUD/HKD). A higher real exchange rate means that the relative purchasing power of your AUD-denominated income is higher."

Statement 4: "All else held equal, a decrease in the nominal exchange rate (AUD/HKD) will decrease the real exchange rate (AUD/HKD) and increase the relative purchasing power of your AUD-denominated income."

To demonstrate the effects of the changes in inflation and nominal exchange rates on relative purchasing power, the adviser uses the following scenario: "Suppose that the AUD/HKD exchange rate increases by 5 percent, the prices of goods and services in Hong Kong go up by 5 percent, and the prices of Australian goods and services go up by 2 percent."

1. Statement 1 is:
   A. correct.
   B. incorrect, because based on the quote convention the investment's value would be decreasing in AUD terms.
   C. incorrect, because the nominal AUD value of the foreign investments will depend on movements in the Australian inflation rate.
2. Statement 2 is:
   A. correct.
   B. incorrect, because purchasing power is not affected in this case.
   C. incorrect, because based on the quote convention, the client's relative purchasing power would be decreasing.
3. Statement 3 is:
   A. correct.
   B. incorrect with respect to the real exchange rate only.
   C. incorrect with respect to both the real exchange rate and the purchasing power of AUD-denominated income.
4. Statement 4 is:
   A. correct.
   B. incorrect with respect to the real exchange rate.
   C. incorrect with respect to the purchasing power of AUD-denominated income.

5. Based on the adviser's scenario and assuming that the HKD value of the HKD bonds remained unchanged, the nominal AUD value of the client's HKD investments would:
   A. decrease by about 5 percent.
   B. increase by about 5 percent.
   C. remain approximately the same.
6. Based on the adviser's scenario, the change in the relative purchasing power of the client's AUD-denominated income is *closest* to:
   A. −8 percent.
   B. +8 percent.
   C. +12 percent.

Solution to 1: A is correct. Given the quoting convention, an increase in the AUD/HKD rate means that the base currency (HKD) is appreciating (one HKD will buy more AUD). This is increasing the nominal value of the HKD-denominated investments when measured in AUD terms.

Solution to 2: B is the correct choice. When paying for HKD-denominated expenses with HKD-denominated income, the value of the AUD/HKD spot exchange rate (or any other spot exchange rate) would not be relevant. In fact, this is a basic principle of currency risk management: reducing FX risk exposures by denominating assets and liabilities (or income and expenses) in the same currency.

Solution to 3: C is the correct choice. An increase in the Australian (i.e., domestic) inflation rate means that the real exchange rate (measured in domestic/foreign, or AUD/HKD, terms) would be decreasing, not increasing. Moreover, an increase in the real exchange rate ($R_{AUD/HKD}$) would be equivalent to a reduction of the purchasing power of the Australian client: Goods and services denominated in HKD would cost more.

Solution to 4: A is correct. As the spot AUD/HKD exchange rate decreases, the HKD is depreciating against the AUD; or equivalently, the AUD is appreciating against the HKD. This is reducing the real exchange rate ($R_{AUD/HKD}$) and increasing the Australian client's purchasing power.

Solution to 5: B is correct. As the AUD/HKD spot exchange rate increases by 5 percent, the HKD is appreciating against the AUD by 5 percent and, all else being equal, the value of the HKD-denominated investment is increasing by 5 percent in AUD terms.

Solution to 6: A is correct. The real exchange rate ($R_{AUD/HKD}$) is expressed as:

$$R_{\frac{AUD}{HKD}} = S_{\frac{AUD}{HKD}} \times \frac{P_{HKD}}{P_{AUD}}$$

The information in the adviser's scenario can be expressed as:

$$\%\Delta R_{\frac{AUD}{HKD}} \approx \%\Delta S_{\frac{AUD}{HKD}} + \%\Delta P_{HKD} - \%\Delta P_{AUD} \approx +5\% + 5\% - 2\% \approx +8\%$$

Because the real exchange rate (expressed in AUD/HKD terms) has gone up by about 8 percent, the real purchasing power of the investor based in Australia has declined by about 8 percent. This can be seen from the fact that HKD has appreciated against the AUD in nominal terms, and the Hong Kong price level has also increased. This increase in the cost of Hong Kong goods and services (measured in AUD) is only partially offset by the small (2 percent) increase in the investor's income (assumed equal to the change in the Australian price level).

2.1. Market Functions

Foreign exchange markets facilitate international trade in goods and services, where companies and individuals need to make transactions in foreign currencies. This would cover everything from companies and governments buying and selling products in other countries, to tourists engaged in cross-border travel (for example, a German tourist selling euros and buying sterling for a visit to London). Although this is an important dimension of FX markets, and despite the growth of global trade in recent years, an even larger proportion of the daily turnover in FX markets is accounted for by capital market transactions, where investors convert between currencies for the purpose of moving funds into (or out of) foreign assets. These types of transactions cover the range from direct investments (for example, companies buying such fixed assets as factories) in other countries to portfolio investments (the purchase of stocks, bonds, and other financial assets denominated in foreign currencies). Because capital is extremely mobile in modern financial markets, this ebb and flow of money across international borders and currencies generates a huge and growing volume of FX transactions.

Regardless of the underlying motivation for the FX transaction, it will eventually require that one currency be exchanged for another in the FX market. In advance of that required transaction, market participants are exposed to the risk that the exchange rate will move against them. Often they will try to reduce (hedge) this risk through a variety of FX instruments (described in more detail later). Conversely, market participants may form opinions about future FX movements and undertake speculative FX risk exposures through a variety of FX instruments in order to profit from their views.

The distinction between hedging and speculative positions is not always clear-cut. For example, consider the case of a corporation selling its products overseas. This creates an FX risk exposure because the revenue from foreign sales will ultimately need to be converted into the corporation's home currency. This risk exposure is typically hedged, and corporate hedging often accounts for large FX flows passing through the market. The amount and timing of foreign revenue, however, are generally hard to predict with precision: they will depend on the pace of foreign sales, the sales prices realized, the pace at which foreign clients pay for their purchases, and so forth. In the face of this uncertainty, the corporate treasury will

estimate the timing and amount of foreign revenue and will then hedge a portion of this estimated amount. Many corporate treasuries have hedging targets based on this estimate, but they also have the flexibility to underhedge or overhedge based on their opinions about future FX rate movements. In order to judge the effectiveness of these discretionary trades, the performance of the corporate treasury is compared with a benchmark, usually stated in terms of a fixed amount hedged relative to total sales. (For example, the benchmark may be a 100 percent fully hedged position. The profitability of the hedge actually implemented—which, based on the treasury's discretion, can vary above or below 100 percent—is then compared with what would have been achieved with a passive, 100 percent fully hedged position.) A corporate treasury manager's performance is judged based on gains or losses relative to the benchmark, just as an investment fund manager's performance is benchmarked against performance targets.

At the other end of the spectrum between hedging and speculation, consider the archetypical speculative account: a hedge fund. Although it is true that hedge funds will seek out, accept, and manage risk for profit, a hedge fund is, after all, a hedge fund; strict risk control procedures are critical to the fund's success, especially when leverage is involved. This mixture of speculative and hedging motives is common throughout the FX space as market participants shape their FX exposures to suit their market forecasts, operational mandates, and appetites for risk.

The FX market provides a variety of products that provide the flexibility to meet this varied and complex set of financial goals. *Spot* transactions involve the exchange of currencies for immediate delivery. For most currencies, this corresponds to "T + 2" delivery, meaning that the exchange of currencies is settled two business days after the trade is agreed to by the two sides of the deal. (One exception is the Canadian dollar, for which spot settlement against the U.S. dollar is on a T + 1 basis.) The exchange rate used for these spot transactions is referred to as the spot exchange rate, and it is the exchange rate that most people refer to in their daily lives (for example, this is the exchange rate usually quoted by the financial press, on the evening news, and so forth).

It is important to realize, however, that spot transactions make up only a minority of total daily turnover in the global FX market; the rest is accounted for by trade in outright forward contracts, FX swaps, and FX options.

Outright *forward contracts* (often referred to simply as forwards) are agreements to deliver foreign exchange at a future date at an exchange rate agreed upon today. For example, suppose that a UK-based company expects to receive a payment of 100 million euros in 85 days. Although it could convert these euros to British pounds with a spot transaction (the spot rate would be the GBP/EUR rate in 83 days, because of T + 2 settlement), this future spot rate is currently unknown and represents a foreign exchange risk to the company. The company can avoid this risk by entering into a transaction with a foreign exchange dealer to sell 100 million euros against the British pound for settlement 85 days from today at a rate—the forward exchange rate—agreed upon today.

Forward contracts are any exchange rate transactions that occur with currency settlement longer than the usual T + 2 settlement for spot delivery. Each of these contracts requires two specifications: the date at which the currencies are to be exchanged and the exchange rate to be applied on the settlement date. Accordingly, exchange rates for these transactions are called *forward exchange rates* to distinguish them from spot rates.

Dealers will typically quote forward rates for a variety of standard forward settlement dates (for example, one week, one month, or 90 days) on their dealing screens. In an over-the-counter (OTC) market, however, traders can arrange forward settlement at *any* future date

they agree on, with the forward exchange rate scaled appropriately for the specific term to settlement. Standard forward settlement dates (such as three months) are defined in terms of the spot settlement date, which is generally T + 2. For example, if today is 18 October and spot settlement is for 20 October, then a three-month forward settlement would be defined as 20 January of the following year. Note as well that these standard forward settlement dates may not always be good business days: 20 January could be a weekend or a holiday. In that case, the forward settlement date is set to the closest good business day. Traders always confirm the exact forward settlement date when making these types of trades, and the forward rate is scaled by the exact number of days to settlement.

In an OTC market, the size of the forward contracts can also be any size that the two counterparties agree on. In general, however, liquidity in forward markets declines the longer the term to maturity and the larger the trade size. The concept of the forward exchange rate and exchange hedging is developed further in Section 3.

Although the OTC market accounts for the majority of foreign exchange trades with future (i.e., greater than T + 2) settlement dates, there is also a deep, liquid market in exchange-traded *futures* contracts for currencies. Although there are technical differences between futures and forward contracts, the basic concept is the same: the price is set today for settlement on a specified future date. Futures contracts on currencies trade on several exchanges globally, but the majority of volume in exchange-traded currency futures contracts is found on the International Monetary Market (IMM) division of the Chicago Mercantile Exchange (CME Group). Futures contracts differ from OTC forward contracts in several important ways: they trade on exchanges (such as the CME) rather than OTC; they are available only for fixed contract amounts and fixed settlement dates; the exchanges demand that a fixed amount of collateral be posted against the futures contract trade; and this collateral is marked to market daily, with counterparties asked to post further collateral if their positions generate losses. On balance, futures contracts are somewhat less flexible than forward contracts. Nonetheless, they provide deep, liquid markets for deferred delivery with a minimum of counterparty (i.e., default) risk—a proposition that many FX traders find attractive. Accordingly, daily turnover in FX futures contracts is huge. As of 2010, the average daily trading volume of FX futures on the CME alone was estimated to be about USD140 billion, which is almost comparable in size to the interbank volume of spot transactions.

Because forward contracts eventually expire, existing speculative positions or FX hedges that need to be extended must be rolled prior to their settlement dates. This typically involves a spot transaction to offset (settle) the expiring forward contract and a new forward contract to be set at a new, more distant settlement date. The combination of an offsetting spot transaction and a new forward contract is referred to as an **FX swap**.[2]

An FX swap is best illustrated by an example. Suppose that a trader sells 100 million euros with settlement 95 days from today at a forward exchange rate (USD/EUR) of 1.2500. In 93 days, the forward contract is two days from settlement, specifically the T + 2 days to spot settlement. To roll the forward contract, the trader will engage in the following FX swap. First, the trader will need to buy 100 million euros spot, for which T + 2 settlement will fall on day 95, the same day as the settlement of the expiring forward contract. The purchase of the 100 million euros spot will be used to satisfy the delivery of the 100 million euros sold in the expiring forward contract. Because 100 million euros are being both bought and sold on day

[2]Note that an FX swap is not the same as a currency swap. An FX swap is simply the combination of a spot and a forward FX transaction (i.e., only two settlement dates—spot and forward—are involved). A currency swap is generally used for multiple periods and payments.

95, there is no exchange of euros between counterparties on that day: The amounts net to zero. However, there will be an exchange of U.S. dollars, reflecting the movement in exchange rates between the date the forward contract was agreed to (day 0) and day 93. Suppose that on day 93 the spot exchange rate for USD/EUR is 1.2400. This means that the trader will see a cash flow on day 95 of USD1,000,000. This is calculated as follows:

$$\text{EUR}100{,}000{,}000 \times (1.2500 - 1.2400) = \text{USD}1{,}000{,}000$$

The trader receives USD1,000,000 from the counterparty because the euro was *sold* forward to day 95 at a price of 1.2500; it was *bought* (on day 93) for spot settlement on day 95 at a price of 1.2400. This *price* movement in the euro indicates a profit to the trader, but because the euro *quantities* exchanged on day 95 net to zero (100,000,000 euros both bought and sold), this cash flow is realized in U.S. dollars. The second leg of the FX swap is then to initiate a new forward sale of 100 million euros at the USD/EUR forward exchange rate being quoted on day 93. This renews the forward position (a forward sale of the euro) to a new date.

For the purposes of this chapter, it is only necessary to understand that (1) an FX swap consists of a simultaneous spot and forward transaction; (2) these swap transactions can extend (roll) an existing forward position to a new future date; and (3) rolling the position forward leads to a cash flow on settlement day. This cash flow can be thought of as a mark-to-market on the forward position. FX swaps are a large component of daily FX market turnover because market participants have to roll over existing speculative or hedging positions as the underlying forward contracts mature in order to extend the hedge or speculative position (otherwise, the position is closed out on the forward settlement date).

One other area where FX swaps are used in FX markets also bears mentioning: They are often used by market participants as a funding source (called swap funding). Consider the case of a UK-based firm that needs to borrow GBP100 million for 90 days, starting two days from today. One way to do this is simply to borrow 90-day money in GBP-denominated funds starting at T + 2. An alternative is to borrow in U.S. dollars and exchange these for British pounds in the spot FX market (both with T + 2 settlement) and then sell British pounds 90 days forward against the U.S. dollar. (Recall that the maturity of a forward rate contract is defined in terms of the spot settlement date, so the 90-day forward rate would be for settlement in 92 days from today.) The company has the use of GBP100 million for 90 days, starting on T + 2, and at the end of this period can pay off the U.S. dollar loan at a known, predetermined exchange rate (the 90-day forward rate). By engaging in simultaneous spot and forward transactions (i.e., an FX swap), the company has eliminated any FX risk from the foreign borrowing. The all-in financing rate using an FX swap will typically be close to that of domestic borrowing, usually within a few basis points. This near equivalence is enforced by an arbitrage relationship that will be described in Section 3.3. On large borrowing amounts, however, even a small differential can add up to substantial cost savings.

Another way to hedge FX exposures, or implement speculative FX positions, is to use options on currencies. FX options are contracts that, for an up-front premium or fee, give the purchaser the right, but not the obligation, to make an FX transaction at some future date at an exchange rate agreed on today (when the contract is agreed to). The holder of an FX option will exercise the option only if it is advantageous to do so—that is, if the agreed-on exchange rate for the FX option contract is better than the FX rate available in the market at option expiry. As such, options are extremely flexible tools for managing FX exposures and account for a large percentage of daily turnover in the FX market.

Another concept to bear in mind is that spot, forward, swap, and option products are typically not used in isolation. Most major market participants manage their FX transactions and FX risk exposures through concurrent spot, forward, swap, and option positions. Taken together, these instruments (the building blocks of the FX market) provide an extremely flexible way for market participants to shape their FX risk exposures to match their operational mandates, risk tolerances, and market opinions. Moreover, FX transactions are often made in conjunction with transactions in other financial markets—such as equities, fixed income, and commodities. These markets have a variety of instruments as well, and market participants jointly tailor their *overall* positions simultaneously using the building blocks of the FX market and these other markets.

EXAMPLE 9-2 Spot and Forward Exchange Rates

The investment adviser based in Sydney, Australia, continues the meeting with the local client who has diversified her domestic bond portfolio by adding investments in fixed-rate, long-term bonds denominated in HKD. Given that the client spends most of the year in Australia, she remains concerned about the foreign exchange risk of her foreign investments and asks the adviser how these might be managed. The investment adviser explains the difference between spot and forward exchange rates and their role in determining foreign exchange risk exposures. The investment adviser suggests the following investment strategy to the client: "You can exchange AUD for HKD in the spot exchange market, invest in a risk-free, one-year HKD-denominated zero coupon bond, and use a one-year forward contract for converting the proceeds back into AUD."

Spot exchange rate (AUD/HKD)	0.1429
One-year HKD interest rate	7.00%
One-year forward exchange rate (AUD/HKD)	0.1402

1. Which of the following statements is *most* correct? Over a one-year horizon, the exchange rate risk of the client's investment in HKD-denominated bonds is determined by uncertainty over:
   A. today's AUD/HKD forward rate.
   B. the AUD/HKD spot rate one year from now.
   C. the AUD/HKD forward rate one year from now.
2. To reduce the exchange rate risk of the Hong Kong investment, the client should:
   A. sell AUD spot.
   B. sell AUD forward.
   C. sell HKD forward.
3. Over a one-year horizon, the investment proposed by the investment adviser is *most* likely:
   A. risk free.
   B. exposed to interest rate risk.
   C. exposed to exchange rate risk.

4. To set up the investment proposed by the adviser, the client would need to:
   A. sell AUD spot; sell a one-year, HKD-denominated bond; and buy AUD forward.
   B. buy AUD spot; buy a one-year, HKD-denominated bond; and sell AUD forward.
   C. sell AUD spot; buy a one-year, HKD-denominated bond; and buy AUD forward.
5. The return (in AUD) on the investment proposed by the investment adviser is *closest* to:
   A. 5 percent.
   B. 6 percent.
   C. 7 percent.

Solution to 1: B is correct. The exchange rate risk (for an unhedged investment) is defined by the uncertainty over future spot rates. In this case, the relevant spot rate is that which would prevail one year from now. Forward rates that would be in effect one year from now would be irrelevant, and the current forward rate is known with certainty.

Solution to 2: C is correct. The Australian-based investor owns HKD-denominated bonds, meaning that she is long HKD exposure. To hedge this exposure, she could enter into a forward contract to sell the HKD against the AUD for future delivery (that is, match a long HKD exposure in the cash market with a short HKD exposure in the derivatives market). The forward rate is established at the time the forward contract is entered into, eliminating any uncertainty about what exchange rate would be used to convert HKD-denominated cash flows back into AUD.

Solution to 3: A is correct. The investment is risk free because the investment is based on a risk-free, one-year, zero coupon, HKD-denominated bond—meaning there is no default or reinvestment risk. The investment will mature in one year at par; there is no interest rate risk. The use of a forward contract to convert the HKD-denominated proceeds back to AUD eliminates any exchange rate risk.

Solution to 4: C is correct. To create the investment, the client needs to convert AUD to HKD in the spot exchange market, invest in (buy) the one-year HKD bond, and sell the HKD forward/buy the AUD forward. Note that this process is directly comparable to the swap financing approach described in this section of the chapter.

Solution to 5: A is correct. Converting one AUD to HKD in the spot market gives the client $1/0.1429 = \text{HKD}7.00$. Investing this for one year leads to $7.00 \times 1.07 = \text{HKD}7.49$. Selling this amount of HKD at the forward rate gives $7.49 \times 0.1402 = \text{AUD}1.05$ (rounding to two decimal places). This implies an AUD-denominated return of 5 percent.

2.2. Market Participants

We now turn to the counterparties that participate in FX markets. As mentioned previously, there is an extremely diverse range of market participants, ranging in size from multibillion-dollar investment funds down to individuals trading for their own accounts (including foreign tourists exchanging currencies at airport kiosks).

To understand the various market participants, it is useful to separate them into broad categories. One broad distinction is between what the market refers to as the *buy side* and the *sell side*. The sell side generally consists of large FX trading banks (such as Citigroup, UBS, and Deutsche Bank); the buy side consists of clients who use these banks to undertake FX transactions (i.e., buy FX products) from the sell-side banks.

The buy side can be further broken down into several categories:

- *Corporate accounts.* Corporations of all sizes undertake FX transactions during cross-border purchases and sales of goods and services. Many of their FX flows can also be related to cross-border investment flows—such as international mergers and acquisitions (M&A) transactions, investment of corporate funds in foreign assets, and foreign currency borrowing.
- *Real money accounts.* These are investment funds managed by insurance companies, mutual funds, pension funds, endowments, exchange-traded funds (ETFs), and other institutional investors. These accounts are referred to as real money because they are usually restricted in their use of leverage or financial derivatives. This distinguishes them from leveraged accounts (discussed next); however, many institutional investors often engage in some form of leverage, either directly through some use of borrowed funds or indirectly using financial derivatives.
- *Leveraged accounts.* This category, often referred to as the professional trading community, consists of hedge funds, proprietary trading shops, commodity trading advisers (CTAs), high-frequency algorithmic traders, and the proprietary trading desks at banks—and indeed almost any active trading account that accepts and manages FX risk for profit. The professional trading community accounts for a large and growing proportion of daily FX market turnover. These active trading accounts also have a wide diversity of trading styles. Some are macro-hedge funds that take longer-term FX positions based on their views of the underlying economic fundamentals of a currency. Others are high-frequency algorithmic traders that use technical trading strategies (such as those based on moving averages or Fibonacci levels) and whose trading cycles and investment horizons are sometimes measured in milliseconds.
- *Retail accounts.* The simplest example of a retail account is the archetypical foreign tourist exchanging currency at an airport kiosk. However, it is important to realize that as electronic trading technology has reduced the barriers to entry into FX markets and the costs of FX trading, there has been a huge surge in speculative trading activity by retail accounts—consisting of individuals trading for their own accounts as well as smaller hedge funds and other active traders. This also includes households using electronic trading technology to move their savings into foreign currencies (this is relatively widespread among households in Japan, for example). It is estimated that retail trading accounts for as much as 10 percent of all spot transactions in some currency pairs and that this proportion is growing.
- *Governments.* Public entities of all types often have FX needs, ranging from relatively small (e.g., maintaining consulates in foreign countries) to large (e.g., military equipment purchases or maintaining overseas military bases). Sometimes these flows are purely transactional—the business simply needs to be done—and sometimes government FX flows reflect, at least in part, the public policy goals of the government. Some government FX business resembles that of investment funds, although sometimes with a public policy mandate as well. In some countries, public-sector pension plans and public insurance schemes are run by a branch of the government. One example is the Caisse de Dépôt et Placement du Québec, which was created by the Québec provincial government in Canada to manage that province's public-sector pension plans. The Caisse, as it is called, is a relatively large player in financial markets, with about CAD200 billion of assets under

management as of 2010. Although it has a mandate to invest these assets for optimal return, it is also called upon to help promote the economic development of Québec. It should be noted that many governments—both at the federal and at the provincial/state levels—issue debt in foreign currencies; this, too, creates FX flows. Such supranational agencies as the World Bank and the African Development Bank issue debt in a variety of currencies as well.

- *Central banks.* These entities sometimes intervene in FX markets in order to influence either the level or the trend in the domestic exchange rate. This often occurs when the central banks judge their domestic currency to be too weak and when the exchange rate has overshot any concept of equilibrium level (e.g., because of a speculative attack) to the degree that the exchange rate no longer reflects underlying economic fundamentals. Alternatively, central banks also intervene when the FX market has become so erratic and dysfunctional that end users such as corporations can no longer transact necessary FX business. Conversely, sometimes central banks intervene when they believe that their domestic currency has become too strong to the point that it undercuts that country's export competitiveness. The Bank of Japan intervened against yen strength versus the U.S. dollar in 2004 and again in September 2010. Similarly, in 2010 the Swiss National Bank intervened against strength in the Swiss franc versus the euro by selling the Swiss franc on the CHF/EUR cross-rate. Central bank reserve managers are also frequent participants in FX markets in order to manage their countries' FX reserves. In this context, they act much like real money investment funds—although generally with a cautious, conservative mandate to safeguard the value of their country's foreign exchange reserves. The foreign exchange reserves of some countries are enormous (e.g., China has about USD2.5 trillion in reserves as of 2010), and central bank participation in FX markets can sometimes have a material impact on exchange rates even when these reserve managers are not intervening for public policy purposes. Exhibit 9-2 provides information on central bank reserve holdings as of the second quarter of 2010.[3]

Note that the amount of foreign exchange reserves now held by emerging economies comfortably exceeds those held by developed economies. This largely reflects the rapid growth in foreign reserves held by Asian central banks, because these countries typically run large current account surpluses with the United States and other developed economies. Reserve accumulation by energy-exporting countries in the Middle East and elsewhere is also a factor. Most of the global currency reserves are held in U.S. dollars; the percentage held in USD is more than twice the portion held in the euro, the second most widely held currency in central bank foreign exchange reserves.

EXHIBIT 9-2 Currency Composition of Official Foreign Exchange Reserves, as of 2nd Quarter 2010 (USD billions)

Total foreign exchange holdings globally	8,422
Held by advanced economies	2,927
Held by emerging and developing economies	5,495
Percent of global holdings held in the U.S. dollar[a]	62%

[a]This percentage is calculated using that amount of global currency reserves for which the currency composition can be identified.

[3]See International Monetary Fund (2010b).

- *Sovereign wealth funds (SWFs).* Many countries with large current account surpluses have diverted some of the resultant international capital flows into SWFs rather than into foreign exchange reserves managed by central banks. Although SWFs are government entities, their mandate is usually more oriented to purely investment purposes rather than public policy purposes. As such, SWFs can be thought of as akin to real money accounts, although some SWFs can employ derivatives or engage in aggressive trading strategies. It is generally understood that SWFs use their resources to help fulfill the public policy mandate of their government owners. The SWFs of many current account surplus countries (such as exporting countries in East Asia or oil-exporting countries) are enormous, and their FX flows can be an important determinant of exchange rate movements in almost all of the major currency pairs.

As mentioned, the sell side generally consists of the FX dealing banks that sell FX products to the buy side. Even here, however, distinctions can be made.

- A large and growing proportion of the daily FX turnover is accounted for by the very largest dealing banks, such as Deutsche Bank, Citigroup, UBS, HSBC, and a few other multinational banking behemoths. Maintaining a competitive advantage in FX requires huge fixed-cost investments in the electronic technology that connects the FX market, and it also requires a broad, global client base. As a result, only the largest banks are able to compete successfully in providing competitive price quotes to clients across the broad range of FX products. In fact, among the largest FX dealing banks, a large proportion of their business is crossed internally, meaning that these banks are able to connect buyers and sellers within their own extremely diverse client base and have no need to show these FX flows outside of the bank.
- All other banks fall into the second and third tiers of the FX market sell side. Many of these financial institutions are regional or local banks with well-developed business relationships, but they lack the economies of scale, broad global client bases, or information technology (IT) expertise required to offer competitive pricing across a wide range of currencies and FX products. In many cases, these are banks in emerging markets that don't have the business connections or credit lines required to access the FX market on a cost-effective basis on their own. As a result, these banks often outsource FX services by forming business relationships with the larger first-tier banks; otherwise, they depend on the deep, competitive liquidity provided by the largest FX market participants.

All of the categories and subcategories just listed, on both the buy side and the sell side, are relatively loose, and there is substantial blurring among these groups. For example, in the past, FX price quotes were provided exclusively by sell-side banks; more recently, however, hedge funds and other large players are accessing the professional FX market on equal terms with the dealing banks. Often it is these more aggressive players that act as market makers rather than the banks themselves; banks sometimes depend on liquidity supplied by accounts that might otherwise be described as buy side.

One of the most important ideas to draw from this categorization of market participants is that there is an extremely wide variety of FX market participants, reflecting a complex mix of trading motives and strategies that can vary with time. Most market participants reflect a combination of hedging and speculative motives in tailoring their FX risk exposures. Among public-sector market participants, public policy motives may also be a factor. The dynamic, complex interaction of FX market participants and their trading objectives makes it difficult to analyze or predict movements in FX rates with any precision, or to describe the FX market adequately with simple characterizations.

2.3. Market Size and Composition

In this section, we present a descriptive overview of the global FX market drawn from the 2010 Triennial Survey undertaken by the Bank for International Settlements (2010). The BIS is an umbrella organization for the world's central banks. Every three years, participating central banks undertake a survey of the FX market in their jurisdictions, the results of which are aggregated and compiled at the BIS. The most recent survey, taken in April 2010, gives a broad indication of the current size and distribution of global FX market flows.

As of April 2010, the BIS estimates that average daily turnover in the traditional FX market (comprised of spot, outright forward, and FX swap transactions) totaled approximately USD3.9 trillion. Exhibit 9-3 shows the approximate percentage allocation among FX product types, including both traditional FX products and exchange-traded FX derivatives. Note that this table of percentage allocations adds exchange-traded derivatives to the BIS estimate of average daily turnover of USD3.9 trillion; the spot and OTC forwards categories include only transactions that are not executed as part of a swap transaction.

The survey also provides a percentage breakdown of the average daily flows between sell-side banks (called the interbank market); between banks and financial customers (all nonbank financial entities, such as real money and leveraged accounts, SWFs, and central banks); and between banks and nonfinancial customers (such as corporations, retail accounts, and governments). The breakdown is provided in Exhibit 9-4. It bears noting that the proportion of average daily FX flow accounted for by financial clients is much larger than that for nonfinancial clients. The BIS also reports that the proportion of financial client flows has been growing rapidly, and as of 2010, it exceeded interbank trading volume for the first time. This underscores the fact that only a minority of the daily FX flow is accounted for by corporations and individuals buying and selling foreign goods and services. Huge investment pools and professional traders are accounting for a large and growing proportion of the FX business.

The 2010 BIS survey also identifies the top five currency pairs in terms of their percentage share of average daily global FX turnover. These are shown in Exhibit 9-5. Note that each of these most active pairs includes the U.S. dollar (USD).

EXHIBIT 9-3 FX Turnover by Instrument

Spot	36%
OTC forwards	12
Exchange-traded derivatives	4
Swaps[a]	44
OTC options[b]	4
Total	**100%**

[a]Includes both FX and currency swaps.
[b]Includes what the BIS categorizes as "other FX products."

EXHIBIT 9-4 FX Flows by Counterparty

Interbank	39%
Financial clients	48
Nonfinancial clients	13

EXHIBIT 9-5 FX Turnover by Currency Pair

USD/EUR	28%
JPY/USD	14
USD/GBP	9
USD/AUD	6
CAD/USD	5

The largest proportion of global FX trading occurs in London, followed by New York. This means that FX markets are most active between approximately 8:00 a.m. and 11:30 a.m. New York time, when banks in both cities are open. (The official London close is at 11:00 a.m. New York time, but London markets remain relatively active for a period after that.) Tokyo is the third-largest FX trading hub.

EXAMPLE 9-3 Market Participants and Composition of Trades

The investment adviser based in Sydney, Australia, makes the following statements to a client when describing some of the basic characteristics of the foreign exchange market:

Statement 1: "Foreign exchange transactions for spot settlement see the most trading volume in terms of average daily turnover because the FX market is primarily focused on settling international trade flows."

Statement 2: "The most important foreign exchange market participants on the buy side are corporations engaged in international trade; on the sell side they are the local banks that service their FX needs."

1. Statement 1 is:
   A. correct.
   B. incorrect with respect to the importance of spot settlements.
   C. incorrect with respect to the importance of both spot settlements and international trade flows.
2. Statement 2 is:
   A. correct.
   B. incorrect with respect to corporations engaged in international trade.
   C. incorrect with respect to both corporations and the local banks that service their trade needs.

Solution to 1: C is the correct choice. Although the media generally focus on the spot market when discussing foreign exchange, the majority of average daily trade volume

involves the FX swap market as market participants either roll over or modify their existing hedging and speculative positions (or engage in FX swap financing). Although it is true that all international trade transactions eventually result in some form of spot settlement, this typically generates a great deal of hedging (and speculative) activity in advance of spot settlement. Moreover, an important group of FX market participants engages in purely speculative positioning with no intention of ever delivering/receiving the principal amount of the trades. Most FX trading volume is not related to international trade; portfolio flows (cross-border capital movements) and speculative activities dominate.

Solution to 2: C is the correct choice. As of 2010, the most important foreign exchange market participants in terms of average daily turnover are found not among corporations engaged in international trade but among huge investment managers, both private (e.g., pension funds) and public (e.g., central bank reserve managers or sovereign wealth funds). A large and growing amount of daily turnover is also being generated by high-frequency traders who use computer algorithms to automatically execute extremely high numbers of speculative trades (although their individual ticket sizes are generally small, they add up to large aggregate flows). On the sell side, the largest money center banks (e.g., Deutsche Bank, Citigroup, HSBC, UBS) are increasingly dominating the amount of trading activity routed through dealers. Regional and local banks are increasingly being marginalized in terms of their share of average daily turnover in FX markets.

3. CURRENCY EXCHANGE RATE CALCULATIONS

3.1. Exchange Rate Quotations

Exchange rates represent the relative price of one currency in terms of another. This price can be represented in two ways: (1) currency A buys how many units of currency B or (2) currency B buys how many units of currency A. Of course, these two prices are simply the inverse of each other.

To distinguish between these two prices, market participants sometimes distinguish between *direct* and *indirect* exchange rates. In the quoting convention A/B (where there is a certain number of units of currency A per one unit of currency B), we refer to currency A as the *price currency* (or quote currency); currency B is referred to as the *base currency*. (The reason for this choice of names will become clearer later.) The base currency is always set at a quantity of one. A *direct* currency quote takes the domestic country as the price currency and the foreign country as the base currency. For example, for a Paris-based trader, the domestic currency would be the euro (EUR) and a foreign currency would be the UK pound (GBP). For this Paris-based trader, a *direct* quote would be EUR/GBP. An exchange rate quote of EUR/GBP = 1.2225 means that 1 GBP costs 1.2225 EUR. For this Paris-based trader, an *indirect* quote has the domestic currency—the euro—as the base currency. An indirect quote of GBP/EUR = 0.8180 means that 1 EUR costs 0.8180 GBP. *Direct and indirect quotes are just the inverse (reciprocal) of each other.*

It can be confusing to describe exchange rates as being either direct or indirect because determining the domestic currency and the foreign currency depends on where one is located.

EXHIBIT 9-6 Exchange Rate Quote Conventions

FX Rate Quote Convention	Name Convention	Actual Ratio (Price Currency/ Base Currency)
EUR	Euro	USD/EUR
JPY	Dollar–yen	JPY/USD
GBP	Sterling	USD/GBP
CAD	Dollar–Canada	CAD/USD
AUD	Aussie	USD/AUD
NZD	Kiwi	USD/NZD
CHF	Swiss franc	CHF/USD
EURJPY	Euro–yen	JPY/EUR
EURGBP	Euro–sterling	GBP/EUR
EURCHF	Euro–Swiss	CHF/EUR
GBPJPY	Sterling–yen	JPY/GBP
EURCAD	Euro–Canada	CAD/EUR
CADJPY	Canada–yen	JPY/CAD

For a London-based market participant, the UK pound (GBP) is the domestic currency and the euro (EUR) is a foreign currency. For a Paris-based market participant, it would be the other way around.

To avoid confusion, the professional FX market has developed a set of market conventions that all market participants typically adhere to when making and asking for FX quotes. Exhibit 9-6 displays some of these for the major currencies: the currency code used for obtaining exchange rate quotes, how the market lingo refers to this currency pair, and the actual ratio—price currency per unit of base currency—represented by the quote.

Several things should be noted in this exhibit. First, the three-letter currency codes in the first column (for FX rate quotes) refer to what are considered the major exchange rates. Remember that an exchange rate is the price of one currency in terms of another: There are always two currencies involved in the price. This is different from referring to a single currency in its own right. For example, one can refer to the euro (EUR) as a *currency*; but if we refer to a euro *exchange rate* (EUR), it is always the price of the euro in terms of another currency, in this case the U.S. dollar. This is because in the professional FX market, the three-letter code EUR is always taken to refer to the euro–U.S. dollar exchange rate, which is quoted in terms of the number of U.S. dollars per euro (USD/EUR). Second, where there are six-letter currency codes in the first column, these refer to some of the major *cross-rates*. This topic will be covered in the next section, but generally these are secondary exchange rates and they are not as common as the main exchange rates. (It can be noted that three-letter codes are always in terms of an exchange rate involving the U.S. dollar, whereas the six-letter codes are not.) Third, when both currencies are mentioned in the code or the name convention, *the base currency is always mentioned first, the opposite order of the actual ratio (price currency/base currency).* Thus, the name convention code for "sterling–yen" is "GBPJPY," but the actual

number quoted is the number of yen per sterling (JPY/GBP). It should also be noted that *the codes may appear in a variety of formats that all mean the same thing.* For example, GBPJPY might instead appear as GBP:JPY or GBP–JPY. Fourth, regardless of where a market participant is located, there is always a mix of direct and indirect quotes in common market usage. For example, a trader based in Toronto will typically refer to the euro–Canada and Canada–yen exchange rates—a mixture of direct (CAD/EUR) and indirect (JPY/CAD) quotes for a Canadian-based trader. There is no overall consistency in this mixture of direct and indirect quoting conventions in the professional FX market; a market participant just has to become familiar with how the conventions are used.[4]

Another concept involving exchange rate quotes in professional FX markets is that of a *two-sided price*. When a client asks a bank for an exchange rate quote, the bank will provide a *bid* (the price at which the bank is willing to buy the currency) and an *offer* (the price at which the bank is willing to sell the currency). But there are *two* currencies involved in an exchange rate quote, which is always the price of one currency relative to the other. So, which one is being bought and sold in this two-sided price quote? This is where the lingo involving the price currency (or quote currency) and the base currency, explained earlier, becomes useful. *The two-sided price quoted by the dealer is in terms of buying or selling the base currency*. It shows the number of units of the price currency that the client will receive from the dealer for one unit of the base currency (the bid) and the number of units of the price currency that the client must sell to the dealer to obtain one unit of the base currency (the offer). Consider the case of a client that is interested in a transaction involving the Swiss franc (CHF) and the euro (EUR). As we have seen, the market convention is to quote this as euro–Swiss (CHF/EUR). The EUR is the base currency, and the two-sided quote (price) shows the number of units of the price currency (CHF) that must be paid or will be received for one euro. For example, a two-sided price in euro–Swiss (CHF/EUR) might look like 1.3405–1.3407. The client will receive CHF1.3405 for selling EUR1 to the dealer and must pay CHF1.3407 to the dealer to buy EUR1. Note that *the price is shown in terms of the price currency* and that *the bid is always less than the offer*: The bank buys the base currency (EUR, in this case) at the low price and sells the base currency at the high price. Buying low and selling high is profitable for banks, and spreading clients—trying to widen the bid/offer spread—is how dealers try to increase their profit margins. However, it should be noted that the electronic dealing systems currently used in professional FX markets are extremely efficient in connecting buyers and sellers globally. Moreover, this worldwide competition for business has compressed most bid/offer spreads to very tight levels. For simplicity, in the remainder of this chapter we will focus on exchange rates as a single number (with no bid/offer spread).

One last thing that can be pointed out in exchange rate quoting conventions is that most major spot exchange rates are typically quoted to four decimal places. One exception among the major currencies involves the yen, for which spot exchange rates are usually quoted to two decimal places. (For example, using spot exchange rates from the middle of 2010, a USD/EUR quote would be expressed as 1.2875, while a JPY/EUR quote would be expressed as

[4]In general, however, there is a hierarchy for quoting conventions. For quotes involving the EUR, it serves as the base currency (e.g., GBP/EUR). Next in the priority sequence, for quotes involving the GBP (but not the EUR) it serves as the base currency (e.g., USD/GBP). Finally, for quotes involving the USD (but not the GBP or EUR) it serves as the base currency (e.g., CAD/USD). Exceptions among the major currencies are the AUD and NZD: they serve as the base currency when quoted against the USD (i.e., USD/AUD, USD/NZD).

110.25.) This difference involving the yen comes from the fact that the units of yen per unit of other currencies is typically relatively large already, and hence extending the exchange rate quote to four decimal places is viewed as unnecessary.

Regardless of what quoting convention is used, changes in an exchange rate can be expressed as a percentage appreciation of one currency against the other: One simply has to be careful in identifying which currency is the price currency and which is the base currency. For example, let's suppose the exchange rate for the euro (USD/EUR) increases from 1.2500 to 1.3000. This represents an unannualized percentage change of:

$$\frac{1.3000}{1.2500} - 1 = 4.00\%$$

This represents a 4 percent appreciation in the euro against the U.S. dollar (and not an appreciation of the U.S. dollar against the euro) because the USD/EUR exchange rate is expressed with the dollar as the price currency.

Note that this appreciation of the euro against the U.S. dollar can also be expressed as a depreciation of the U.S. dollar against the euro; but in this case, the depreciation is not equal to 4.0 percent. Inverting the exchange rate quote from USD/EUR to EUR/USD, so that the euro is now the price currency, leads to:

$$\frac{\left(\frac{1}{1.3000}\right)}{\left(\frac{1}{1.2500}\right)} - 1 = \frac{1.2500}{1.3000} - 1 = -3.85\%$$

Note that the U.S. dollar depreciation is not the same, in percentage terms, as the euro appreciation. This will always be true; it is simply a matter of arithmetic.

EXAMPLE 9-4 Exchange Rate Conventions

A dealer based in New York City provides a spot exchange rate quote of 12.4035 MXN/USD to a client in Mexico City. The inverse of 12.4035 is 0.0806.

1. From the perspective of the Mexican client, the *most* accurate statement is that:
   A. the direct exchange rate quotation is equal to 0.0806.
   B. the direct exchange rate quotation is equal to 12.4035.
   C. the indirect exchange rate quotation is equal to 12.4035.
2. If the bid/offer quote from the dealer was 12.4020–12.4060 MXN/USD, then the bid/offer quote in USD/MXN terms would be *closest* to:
   A. 0.08061–0.08063.
   B. 0.08063–0.08061.
   C. 0.08062–0.08062.

Solution to 1: B is correct. A direct exchange rate uses the domestic currency as the price currency and the foreign currency as the base currency. For an MXN/USD quote, the

MXN is the price currency; therefore, the direct quote for the Mexican client is 12.4035 (it costs 12.4035 pesos to purchase one U.S. dollar). Another way of understanding a *direct* exchange rate quote is that it is the price of one unit of foreign currency in terms of your own currency. This purchase of a unit of foreign currency can be thought of as a purchase much like any other you might make; think of the unit of foreign currency as just another item that you might be purchasing with your domestic currency. For example, for someone based in Canada, a liter of milk currently costs about CAD1.25, and USD1 costs about CAD1.03. This *direct* currency quote uses the *domestic* currency (the Canadian dollar, in this case) as the *price* currency and simply gives the price of a unit of foreign currency that is being purchased.

Solution to 2: A is correct. An MXN/USD quote means the amount of MXN the dealer is bidding to buy USD1 or offering to sell USD1. The dealer's bid to buy USD1 at MXN12.4020 is equivalent to the dealer paying MXN12.4020 to buy USD1. Dividing both terms by 12.4020 means the dealer is paying (i.e., selling) MXN1 to buy USD0.08063. This is the offer in USD/MXN terms: The dealer offers to sell MXN1 at a price of USD0.08063. In USD/MXN terms, the dealer's bid for MXN1 is 0.08061, calculated by inverting the offer of 12.4060 in MXN/USD terms (1/12.4060 = 0.08061). Note that in any bid/offer quote, no matter what the base or price currencies, the bid is always lower than the offer.

3.2. Cross-Rate Calculations

Given two exchange rates involving three currencies, it is possible to back out what the cross-rate must be. For example, as we have seen, the FX market convention is to quote the exchange rate between the U.S. dollar and the euro as euro–dollar (USD/EUR). The FX market also quotes the exchange rate between the Canadian dollar and U.S. dollar as dollar–Canada (CAD/USD). Given these two exchange rates, it is possible to back out the cross-rate between the euro and the Canadian dollar, which according to market convention is quoted as euro–Canada (CAD/EUR). This calculation is shown as:

$$\frac{\text{CAD}}{\text{USD}} \times \frac{\text{USD}}{\text{EUR}} = \frac{\text{CAD}}{\cancel{\text{USD}}} \times \frac{\cancel{\text{USD}}}{\text{EUR}} = \frac{\text{CAD}}{\text{EUR}}$$

Hence, to get a euro–Canada (CAD/EUR) quote, we must multiply the dollar–Canada (CAD/USD) quote by the euro–dollar (USD/EUR) quote. For example, assume the exchange rate for dollar–Canada is 1.0460 and the exchange rate for euro–dollar is 1.2880. Using these sample spot exchange rates, calculating the euro–Canada cross-rate equals:

$$1.0460 \times 1.2880 = 1.3472 \text{ CAD per EUR}$$

It is best to avoid talking in terms of direct or indirect quotes because, as noted earlier, these conventions depend on where one is located and hence what the domestic and foreign

currencies are. Instead, focus on how the math works: Sometimes it is necessary to invert one of the quotes in order to get the intermediary currency to cancel out in the equation to get the cross-rate. For example, to get a Canada–yen (JPY/CAD) quote, one is typically using the dollar–Canada (CAD/USD) rate and dollar–yen (JPY/USD) rate, which are the market conventions. This Canada–yen calculation requires that the dollar–Canada rate (CAD/USD) be inverted to a USD/CAD quote for the calculations to work, as shown:

$$\left(\frac{\text{CAD}}{\text{USD}}\right)^{-1} \times \frac{\text{JPY}}{\text{USD}} = \frac{\text{USD}}{\text{CAD}} \times \frac{\text{JPY}}{\cancel{\text{USD}}} = \frac{\cancel{\text{USD}}}{\text{CAD}} \times \frac{\text{JPY}}{\text{USD}} = \frac{\text{JPY}}{\text{CAD}}$$

Hence, to get a Canada–yen (JPY/CAD) quote, we must first invert the dollar–Canada (CAD/USD) quote before multiplying by the dollar–yen (JPY/USD) quote. As an example, let's assume that we have spot exchange rates of 1.0460 for dollar–Canada (CAD/USD) and 85.50 for dollar–yen (JPY/USD). The dollar–Canada rate of 1.0460 inverts to 0.9560; multiplying this value by the dollar–yen quote of 85.50 gives a Canada–yen quote of:

$$0.9560 \times 85.50 = 81.74 \text{ JPY per CAD}$$

Market participants asking for a quote in a cross-rate currency pair typically will not have to do this calculation themselves; either the dealer or the electronic trading platform will provide a quote in the specified currency pair. (For example, a client asking for a quote in Canada–yen will receive that quote from the dealer; the client will not be given separate dollar–Canada and dollar–yen quotes in order to do the math.) But be aware that dealers providing the quotes may have to do this calculation themselves if only because the dollar–Canada and dollar–yen currency pairs often trade on different trading desks and involve different traders. Electronic dealing machines used in both the inter-bank market and bank-to-client markets often provide this mathematical operation to calculate cross-rates automatically.

Because market participants can receive both a cross-rate quote (for example, Canada–yen) as well as the components underlying exchange rate quotes (for example, dollar–Canada and dollar–yen), these cross-rate quotes must be consistent with the above equation; otherwise, the market will arbitrage the mispricing. Extending our example, we calculate a Canada–yen (JPY/CAD) rate of 81.74 based on underlying dollar–Canada (CAD/USD) and dollar–yen (JPY/USD) rates of 1.0460 and 85.50, respectively. Now suppose that at the same time a misguided dealer quotes a Canada–yen rate of 82.00. This is a different price in JPY/CAD for an identical service: converting yen into Canadian dollars. Hence, any trader could buy CAD1 at the lower price of JPY81.74 and then turn around and sell CAD1 at JPY82.00 (recall our earlier discussion of how price and base currencies are defined). The riskless arbitrage profit is JPY0.26 per CAD1. The arbitrage—called **triangular arbitrage** because it involves three currencies—would continue until the price discrepancy was removed.

In reality, however, these discrepancies in cross-rates almost never occur because both human traders and automatic trading algorithms are constantly on alert for any pricing inefficiencies. In practice, and for the purposes of this chapter, we can consider cross-rates as being consistent with their underlying exchange rate quotes and that given any two exchange rates involving three currencies, we can back out the third cross-rate.

EXAMPLE 9-5 Cross Exchange Rates and Percentage Changes

A research report produced by a dealer includes the following exhibit:

	Spot Rate	Expected Spot Rate in One Year
USD/EUR	1.3960	1.3863
CHF/USD	0.9585	0.9551
USD/GBP	1.5850	1.5794

1. The spot CHF/EUR cross-rate is *closest* to:
   A. 0.6866.
   B. 0.7473.
   C. 1.3381.
2. The spot GBP/EUR cross-rate is *closest* to:
   A. 0.8808.
   B. 1.1354.
   C. 2.2127.
3. Based on the exhibit, the euro is expected to appreciate by how much against the U.S. dollar over the next year?
   A. −0.7 percent
   B. +0.7 percent
   C. +1.0 percent
4. Based on the exhibit, the U.S. dollar is expected to appreciate by how much against the British pound over the next year?
   A. +0.6 percent
   B. −0.4 percent
   C. +0.4 percent
5. Over the next year, the Swiss franc is expected to:
   A. depreciate against the GBP.
   B. depreciate against the EUR.
   C. appreciate against the GBP, EUR, and USD.
6. Based on the exhibit, which of the following lists the three currencies from strongest to weakest over the next year?
   A. USD, GBP, EUR
   B. USD, EUR, GBP
   C. EUR, USD, GBP
7. Based on the exhibit, which of the following lists the three currencies in order of appreciating the most to appreciating the least (in percentage terms) against the USD over the next year?
   A. GBP, CHF, EUR
   B. CHF, GBP, EUR
   C. EUR, CHF, GBP

Solution to 1: C is correct:

$$\frac{\text{CHF}}{\text{EUR}} = \frac{\text{USD}}{\text{EUR}} \times \frac{\text{CHF}}{\text{USD}} = 1.3960 \times 0.9585 = 1.3381$$

Solution to 2: A is correct:

$$\frac{\text{GBP}}{\text{EUR}} = \frac{\text{USD}}{\text{EUR}} \times \left(\frac{\text{USD}}{\text{GBP}}\right)^{-1} = \frac{\text{USD}}{\text{EUR}} \times \frac{\text{GBP}}{\text{USD}} = \frac{1.3960}{1.5850} = 0.8808$$

Solution to 3: A is correct. The euro is the base currency in the USD/EUR quote, and the expected decrease in the USD/EUR rate indicates that the EUR is depreciating (in one year it will cost less USD to buy one EUR). Mathematically:

$$\frac{1.3863}{1.3960} - 1 = -0.7\%$$

Solution to 4: C is correct. The GBP is the base currency in the USD/GBP quote, and the expected decrease in the USD/GBP rate means that the GBP is expected to depreciate against the USD. Or equivalently, the USD is expected to appreciate against the GBP. Mathematically:

$$\left(\frac{1.5794}{1.5850}\right)^{-1} - 1 = \frac{1.5850}{1.5794} - 1 = +0.4\%$$

Solution to 5: C is correct: Because the question does not require calculating the magnitude of the appreciation or depreciation, we can work with CHF as either the price currency or the base currency. In this case, it is easiest to use it as the price currency. According to the table, CHF/USD is expected to decline from 0.9585 to 0.9551, so CHF is expected to be stronger (i.e., it should appreciate against the USD). CHF/EUR is currently 1.3381 (see the solution to Question 1) and is expected to be 1.3241 (= 0.9551 × 1.3863), so CHF is expected to appreciate against the EUR. CHF/GBP is currently 1.5192 (= 0.9585 × 1.5850) and is expected to be 1.5085 (= 0.9551 × 1.5794), so CHF is also expected to appreciate against the GBP.

Alternatively, we can derive this answer intuitively. The table shows that the CHF/USD rate is expected to decline: That is, the USD is expected to depreciate against the CHF, or alternatively, the CHF is expected to appreciate against the USD. The table also shows that the USD/EUR and USD/GBP rates are also decreasing, meaning that the EUR and GBP are expected to depreciate against the USD, or alternatively, the USD is expected to appreciate against the EUR and GBP. If the CHF is expected to

appreciate against the USD and the USD is expected to appreciate against both the EUR and the GBP, it follows that the CHF is expected to appreciate against both the EUR and the GBP.

Solution to 6: A is correct. According to the table, USD/EUR is expected to decline from 1.3960 to 1.3863, while USD/GBP is expected to decline from 1.5850 to 1.5794. So, the USD is expected to be stronger than both the EUR and the GBP. GBP/EUR is currently 0.8808 ($= 1.5850^{-1} \times 1.3960$) and is expected to be 0.8777 ($= 1.5794^{-1} \times 1.3863$), so the GBP is expected to be stronger than the EUR.

Solution to 7: B is correct. The USD/EUR rate depreciates by −0.7 percent [= (1.3863/1.3960) – 1], which is the depreciation of the base currency EUR against the USD. The USD/GBP rate declines −0.4 percent [= (1.5794/1.5850) – 1], which is the depreciation of the GBP against the USD. Inverting the CHF/USD rate to a USD/CHF convention shows that the base currency CHF appreciates by +0.4 percent against the USD [= (1.0470/1.0433) – 1].

3.3. Forward Calculations

In professional FX markets, forward exchange rates are typically quoted in terms of points (also sometimes referred to as *pips*). The points on a forward rate quote are simply the difference between the forward exchange rate quote and the spot exchange rate quote, with the points scaled so that they can be related to the last decimal in the spot quote. When the forward rate is higher than the spot rate, the points are positive and the base currency is said to be trading at a *forward premium*. Conversely, if the forward rate is less than the spot rate, the points (forward rate minus spot rate) are negative and the base currency is said to be trading at a *forward discount*. Of course, if the base currency is trading at a forward premium, then the price currency is trading at a forward discount, and vice versa.

This can best be explained by means of an example. At one point during 2010, the spot euro–dollar exchange rate (USD/EUR) was 1.2875 and the one-year forward rate was 1.28485. Hence, the forward rate was trading at a discount to the spot rate (the forward rate was smaller than the spot rate) and the one-year forward points were quoted as −26.5. This −26.5 comes from:

$$1.28485 - 1.2875 = -0.00265$$

Recall that most non-yen exchange rates are quoted to four decimal places, so in this case we would scale up by four decimal places (multiply by 10,000) so that this −0.00265 would be represented as −26.5 points. Notice that the points are scaled to the size of the last digit in the spot exchange rate quote—usually the fourth decimal place. Notice as well that points are typically quoted to one (or more) decimal places, meaning that the forward rate will typically be quoted to five or more decimal places. The exception among the major currencies is the yen, which is typically quoted to two decimal places for spot rates. Here, forward points are scaled up by two decimal places—the last digit in the spot rate quote—by multiplying the difference between forward and spot rates by 100.

Typically, quotes for forward rates are shown as the number of forward points at each maturity.[5] These forward points are also called *swap points* because an FX swap consists of simultaneous spot and forward transactions. In the middle of 2010, a trader would have faced a spot rate and forward points in the euro–dollar (USD/EUR) currency pair similar to those in Exhibit 9-7.

Notice that the absolute number of points generally increases with maturity. This is because the number of points is proportional to the yield differential between the two countries (the Eurozone and the United States, in this case) scaled by the term to maturity. Given the interest rate differential, the longer the term to maturity, the greater the absolute number of forward points. Similarly, given the term to maturity, a wider interest rate differential implies a greater absolute number of forward points. (This will be explained and demonstrated in more detail later in this section.)

To convert any of these quoted forward points into a forward rate, one would divide the number of points by 10,000 (to scale down to the fourth decimal place, the last decimal place in the spot quote) and then add the result to the spot exchange rate quote.[6] For example, using the data in Exhibit 9-7, the three-month forward rate in this case would be:

$$1.2875 + \left(\frac{-5.5}{10{,}000}\right) = 1.2875 - 0.00055 = 1.28695$$

Occasionally, one will see the forward rate or forward points represented as a percentage of the spot rate rather than as an absolute number of points. Continuing our example, the three-month forward rate for USD/EUR can be represented as:

$$\frac{1.28750 - 0.00055}{1.28750} = \left(\frac{1.28695}{1.28750}\right) - 1 = -0.043\%$$

EXHIBIT 9-7 Sample Spot and Forward Quotes

Maturity	Spot Rate or Forward Points
Spot	1.2875
One week	−0.3
One month	−1.1
Three months	−5.5
Six months	−13.3
Twelve months	−26.5

[5]As mentioned earlier, maturity is defined in terms of the time between spot settlement (usually T + 2) and the settlement of the forward contract.

[6]Because the JPY/USD exchange rate is quoted to only two decimal places, forward points for the dollar–yen currency pair are divided by 100.

This shows that either the forward rate or the forward points can be used to calculate the percentage discount (or premium) in the forward market—in this case, −0.043 percent, rounding to three decimal places. To convert a spot quote into a forward quote when the points are shown as a percentage, one simply multiples the spot rate by 1 plus the percentage premium or discount:

$$1.28750 \times (1 - 0.043\%) = 1.28750 \times (1.0 - 0.00043) \approx 1.28695$$

Note that, rounded to the fifth decimal place, this is equal to our previous calculation. However, it is typically the case in professional FX markets that forward rates will be quoted in terms of pips rather than percentages.

We now turn to the relationship among spot rates, forward rates, and interest rates and how their relationship is derived. Forward exchange rates are based on an arbitrage relationship that equates the investment return on two alternative but equivalent investments. Consider the case of an investor with funds to invest. For simplicity, we will assume that there is one unit of the investor's domestic currency to be invested for one period. One alternative is to invest for one period at the domestic risk-free rate (i_d); at the end of the period, the amount of funds held is equal to $(1 + i_d)$. An alternative investment is to convert this one unit of domestic currency to foreign currency using the spot rate of $S_{f/d}$ (number of units of foreign currency per one unit of domestic currency). This can be invested for one period at the foreign risk-free rate; at the end of the period, the investor would have $S_{f/d}\,(1 + i_f)$ units of foreign currency. These funds must then be converted back to the investor's domestic currency. If the exchange rate to be used for this end-of-period conversion was precontracted at the start of the period (i.e., a forward rate was used), it would eliminate any foreign exchange risk from converting at a future, unknown spot rate. Given the assumed exchange rate convention here (foreign/domestic), the investor would obtain $(1/F_{f/d})$ units of the domestic currency for each unit of foreign currency sold forward. Note that this process of converting domestic funds in the spot FX market, investing at the foreign risk-free rate, and then converting back to the domestic currency with a forward rate is identical to the concept of swap financing described in an earlier section of this chapter.

Hence, we have two alternative investments—both risk-free because both are invested at risk-free interest rates and because any foreign exchange risk was eliminated (hedged) by using a forward rate. Because these two investments are equal in risk characteristics, they must have the same return. Bearing in mind that the currency quoting convention is the number of foreign currency units per single domestic unit (*f*/*d*), this relationship can be stated as:

$$(1 + i_d) = S_{f/d}(1 + i_f)\left(\frac{1}{F_{f/d}}\right)$$

This is an arbitrage relationship because it describes two alternative investments (one on either side of the equal sign) that should have equal returns. If they do not, a riskless arbitrage opportunity exists because an investor can sell short the investment with the lower return and invest the funds in the investment with the higher return; the difference between the two returns is pure profit.[7]

[7]It is because of this arbitrage relationship that the all-in financing rate using swap financing is close to the domestic interest rate.

This formula is perhaps the easiest and most intuitive way to remember the formula for the forward rate, because it is based directly on the underlying intuition (the arbitrage relationship of two alternative but equivalent investments, one on either side of the equal sign). Also, the right-hand side of the equation, for the hedged foreign investment alternative, is arranged in proper time sequence: (1) convert domestic to foreign currency, then (2) invest the foreign currency at the foreign interest rate, and finally (3) convert the foreign currency back to the domestic currency.[8]

This arbitrage equation can be rearranged as needs require. For example, to get the formula for the forward rate, the preceding equation can be restated as:

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f}{1 + i_d}\right)$$

Another way of looking at this is, given the spot exchange rate and the domestic and foreign risk-free interest rates, the forward rate is whatever value completes this equation and eliminates any arbitrage opportunity. For example, let's assume that the spot exchange rate ($S_{f/d}$) is 1.6535, the domestic 12-month risk-free rate is 3.50 percent, and the foreign 12-month risk-free rate is 5.00 percent. The 12-month forward rate ($F_{f/d}$) must then be equal to:

$$1.6535\left(\frac{1.0500}{1.0350}\right) = 1.6775$$

Suppose instead that, with the spot exchange rate and interest rates unchanged, you were given a quote on the 12-month forward rate ($F_{f/d}$) of 1.6900. Because this misquoted forward rate does not agree with the arbitrage equation, it would present a riskless arbitrage opportunity. This can be seen by using the arbitrage equation to compute the return on the two alternative investment strategies. The return on the domestic-only investment approach is the domestic risk-free rate (3.50 percent). In contrast, the return on the hedged foreign investment when this misquoted forward rate is put into the arbitrage equation equals:

$$S_{f/d}(1 + i_f)\left(\frac{1}{F_{f/d}}\right) = 1.6535(1.05)\left(\frac{1}{1.6900}\right) = 1.0273$$

This defines a return of 2.73 percent. Hence, the investor could make riskless arbitrage profits by borrowing at the higher foreign risk-free rate, selling the foreign currency at the spot exchange rate, hedging the currency exposure (buying the foreign currency back) at the misquoted forward rate, investing the funds at the lower domestic risk-free rate, and thereby getting a profit of 77 basis points (3.50% – 2.73%) for each unit of domestic currency involved—all with no up-front commitment of the investor's own capital. Any such opportunity in real-world financial markets would be quickly arbitraged ("arbed") away. It is interesting to note that in this example, the investor actually borrows at the higher of the two interest rates but makes a profit because the foreign currency is underpriced in the forward market.

[8]Recall that this equation is based on an *f/d* exchange rate quoting convention. If the exchange rate data were presented in *d/f* form, one could either invert these quotes back to *f/d* form and use the preceding equation or use the following equivalent equation: $(1 + i_d) = (1/S_{d/f})(1 + i_f)F_{d/f}$. If this latter equation were used, remember that forward and spot exchange rates are now being quoted on a *d/f* convention.

The underlying arbitrage equation can also be rearranged to show the forward rate as a percentage of the spot rate:

$$\frac{F_{f/d}}{S_{f/d}} = \left(\frac{1 + i_f}{1 + i_d}\right)$$

This shows that, given an *f/d* quoting convention, the forward rate will be higher than (be at a premium to) the spot rate if foreign interest rates are higher than domestic interest rates. More generally, and regardless of the quoting convention, *the currency with the higher interest rate will always trade at a discount in the forward market, and the currency with the lower interest rate will always trade at a premium in the forward market*.

One context in which forward rates are quoted as a percentage of spot rates occurs when forward rates are interpreted as expected future spot rates, or:

$$F_t = \hat{S}_{t+1}$$

Substituting this expression into the previous equation and doing some rearranging leads to:

$$\frac{\hat{S}_{t+1}}{S_t} - 1 = \%\Delta\hat{S}_{t+1} = \left(\frac{i_f - i_d}{1 + i_d}\right)$$

This shows that if forward rates are interpreted as expected future spot rates, the expected percentage change in the spot rate is proportional to the interest rate differential ($i_f - i_d$).

It is intuitively appealing to see forward rates as expected future spot rates. However, this interpretation of forward rates should be used cautiously. First, the direction of the expected change in spot rates is somewhat counterintuitive. All else being equal, an increase in domestic interest rates (for example, the central bank tightens monetary policy) would typically be expected to lead to an increase in the value of the domestic currency. In contrast, the preceding equation indicates that, all else being equal, a higher domestic interest rate implies slower expected appreciation (or greater expected depreciation) of the domestic currency (recall that this equation is based on an *f/d* quoting convention).

More important, historical data show that forward rates are poor predictors of future spot rates. Although various econometric studies suggest that forward rates may be unbiased predictors of future spot rates (i.e., they do not systematically over- or underestimate future spot rates), this is not particularly useful information because the margin of error for these forecasts is so large. As we have seen in our introductory section, the FX market is far too complex and dynamic to be captured by a single variable, such as the level of the yield differential between countries. Moreover, as can be seen in the preceding formula for the forward rate, forward rates are based on domestic and foreign interest rates. This means that anything that affects the level and shape of the yield curve in either the domestic or the foreign market will also affect the relationship between spot and forward exchange rates. In other words, FX markets do not operate in isolation but will reflect almost all factors affecting other markets globally; anything that affects expectations or risk premiums in these other markets will reverberate in forward exchange rates as well. Although the level of the yield differential is one factor that the market may look at in forming spot exchange rate expectations, it is only one of many factors. (Many traders look to the trend in the yield differential rather than the level of

the differential.) Moreover, there is a lot of noise in FX markets that makes almost any model—no matter how complex—a relatively poor predictor of spot rates at any given point in the future. In practice, FX traders and market strategists do *not* base either their currency expectations or their trading strategies solely on forward rates.

For the purposes of this chapter, it is best to understand forward exchange rates simply as a product of the arbitrage equation outlined earlier and forward points as being related to the (time-scaled) interest rate differential between the two countries. Reading any more than that into forward rates or interpreting them as the market forecast can be potentially misleading.

To understand the relationship between maturity and forward points, we need to generalize our arbitrage formula slightly. Suppose the investment horizon is a fraction, τ, of the period for which the interest rates are quoted. Then the interest earned in the domestic and foreign markets would be $(i_d\tau)$ and $(i_f\tau)$, respectively. Substituting this into our arbitrage relationship and solving for the difference between the forward and spot exchange rates gives:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{i_f - i_d}{1 + i_d\tau}\right)\tau$$

This equation shows that forward points (appropriately scaled) are proportional to the spot exchange rate and to the interest rate differential and approximately (but not exactly) proportional to the horizon of the forward contract.

Let's demonstrate this using an example. Suppose that we wanted to determine the 30-day forward exchange rate given a 30-day domestic risk-free interest rate of 2.00 percent per year, a 30-day foreign risk-free interest rate of 3.00 percent per year, and a spot exchange rate ($S_{f/d}$) of 1.6555. The risk-free assets used in this arbitrage relationship are typically bank deposits quoted using the London Interbank Offered Rate (LIBOR) for the currencies involved. The day count convention for LIBOR deposits is actual/360.[9] Incorporating the fractional period (τ) as previously and inserting the data into the forward rate equation leads to a 30-day forward rate of:

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f\tau}{1 + i_d\tau}\right) = 1.6555\left[\frac{1 + 0.0300\left(\frac{30}{360}\right)}{1 + 0.0200\left(\frac{30}{360}\right)}\right] = 1.6569$$

This means that, for a 30-day term, forward rates are trading at a premium of 14 pips (1.6569 – 1.6555). This can also be calculated using the previous formula for swap points:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{i_f - i_d}{1 + i_d\tau}\right)\tau = 1.6555\left[\frac{0.0300 - 0.0200}{1 + 0.0200\left(\frac{30}{360}\right)}\right]\left(\frac{30}{360}\right) = 0.0014$$

As should be clear from this expression, the absolute number of swap points will be closely related to the term of the forward contract (i.e., approximately proportional to $\tau =$ Actual/360). For example, leaving the spot exchange rate and interest rates unchanged, let's set the term of the forward contract to 180 days:

[9]This means that for interest calculation purposes, it is assumed that there are 360 days in the year. However, the actual number of days the funds are on deposit is used to calculate the interest payable.

$$F_{f/d} - S_{f/d} = 1.6555\left[\frac{0.0300 - 0.0200}{1 + 0.0200\left(\frac{180}{360}\right)}\right]\left(\frac{180}{360}\right) = 0.0082$$

This leads to the forward rate trading at a premium of 82 pips. The increase in the number of forward points is approximately proportional to the increase in the term of the contract (from 30 days to 180 days). Note that although the term of the 180-day forward contract is six times longer than that of a 30-day contract, the number of forward points is not exactly six times larger: $6 \times 14 = 84$.

Similarly, the number of forward points is proportional to the spread between foreign and domestic interest rates ($i_f - i_d$). For example, with reference to the original 30-day forward contract, let's set the foreign interest rate to 4.00 percent, leaving the domestic interest rate and spot exchange rate unchanged. This doubles the interest rate differential ($i_f - i_d$) from 1.00 percent to 2.00 percent; it also doubles the forward points (rounding to four decimal places):

$$F_{f/d} - S_{f/d} = 1.6555\left[\frac{0.0400 - 0.0200}{1 + 0.0200\left(\frac{30}{360}\right)}\right]\left(\frac{30}{360}\right) = 0.0028$$

EXAMPLE 9-6 Forward Rates

A French company has recently finalized a sale of goods to a UK-based client and expects to receive a payment of GBP50 million in 32 days. The corporate treasurer at the French company wants to hedge the foreign exchange risk of this transaction and receives the following exchange rate information from a dealer:

GBP/EUR spot rate	0.8752
One-month forward points	−1.4

1. Given this data, the treasurer could hedge the foreign exchange risk by:
   A. buying EUR (selling GBP) at a forward rate of 0.87380.
   B. buying EUR (selling GBP) at a forward rate of 0.87506.
   C. selling EUR (buying GBP) at a forward rate of 0.87506.
2. The *best* interpretation of the forward discount shown is that:
   A. the euro is expected to depreciate over the next 30 days.
   B. one-month UK interest rates are higher than those in the Eurozone.
   C. one-month Eurozone interest rates are higher than those in the United Kingdom.

3. If the 12-month forward rate is 0.87295 GBP/EUR, then, based on the data, the 12-month forward points are *closest* to:
   A. −22.5.
   B. −2.25.
   C. −0.00225.
4. If a second dealer quotes GBP/EUR at a 12-month forward discount of 0.30 percent on the same spot rate, the French company could:
   A. trade with either dealer because the 12-month forward quotes are equivalent.
   B. lock in a profit in 12 months by buying EUR from the second dealer and selling it to the original dealer.
   C. lock in a profit in 12 months by buying EUR from the original dealer and selling it to the second dealer.
5. If the 270-day LIBOR rates (annualized) for the EUR and GBP are 1.370 percent and 1.325 percent, respectively, and the spot GBP/EUR exchange rate is 0.8489, then the number of forward points for a 270-day forward rate ($F_{GBP/EUR}$) is *closest* to:
   A. −22.8.
   B. −3.8.
   C. −2.8.

Solution to 1: B is correct. The French company would want to convert the GBP to its domestic currency, the EUR (it wants to sell GBP, buy EUR). The forward rate would be equal to 0.8752 + (−1.4/10,000) = 0.87506.

Solution to 2: C is correct. A forward discount indicates that interest rates in the base currency country (France in this case, which uses the euro) are higher than those in the price currency country (the United Kingdom).

Solution to 3: A is correct. The number of forward points is equal to the scaled difference between the forward rate and the spot rate. In this case, 0.87295 – 0.87520 = −0.00225. This is then multiplied by 10,000 to convert to the number of forward points.

Solution to 4: B is correct. A 0.30 percent discount means that the second dealer will sell euros 12 months forward at 0.8752 × (1 – 0.0030) = 0.87257, a lower price per euro than the original dealer's quote of 0.87295. Buying euros at the cheaper 12-month forward rate (0.87257) and selling the same amount of euros 12 months forward at the higher 12-month forward rate (0.87295) means a profit of (0.87295 – 0.87257 = GBP 0.00038) per euro transacted, receivable when both forward contracts settle in 12 months.

Solution to 5: C is correct, because the forward rate is calculated as:

$$F_{\frac{GBP}{EUR}} = S_{\frac{GBP}{EUR}}\left[\frac{1 + i_{GBP}\left(\frac{Actual}{360}\right)}{1 + i_{EUR}\left(\frac{Actual}{360}\right)}\right] = 0.8489\left[\frac{1 + 0.01325\left(\frac{270}{360}\right)}{1. + 0.01370\left(\frac{270}{360}\right)}\right] = 0.84862$$

This shows that the forward points are at a discount of 0.84862 – 0.84890 = −0.00028, or −2.8 points. This can also be seen using the swap points formula:

$$F_{\frac{GBP}{EUR}} - S_{\frac{GBP}{EUR}} = 0.8489\left[\frac{0.01325 - 0.01370}{1 + 0.01370\left(\frac{270}{360}\right)}\right]\left[\frac{270}{360}\right] = -0.00028$$

The calculation of −3.8 points omits the day count (270/360), and −22.8 points gets the scaling wrong.

4. EXCHANGE RATE REGIMES

Highly volatile exchange rates create uncertainty that undermines the efficiency of real economic activity and the financial transactions required to facilitate that activity. Exchange rate volatility also has a direct impact on investment decisions because it is a key component of the risk inherent in foreign (i.e., foreign-currency-denominated) assets. Exchange rate volatility is also a critical factor in selecting hedging strategies for foreign currency exposures.

The amount of foreign exchange rate volatility will depend, at least in part, on the institutional and policy arrangements associated with trade in any given currency. Virtually every exchange rate is managed to some degree by central banks. The policy framework that each central bank adopts is called an *exchange rate regime*. Although there are many potential variations, these regimes fall into a few general categories. Before describing each of these types, we consider the possibility of an ideal regime and provide some historical perspective on the evolution of currency arrangements.

4.1. The Ideal Currency Regime

The ideal currency regime would have three properties. First, the exchange rate between any two currencies would be credibly fixed. This would eliminate currency-related uncertainty with respect to the prices of goods and services as well as real and financial assets. Second, all currencies would be fully convertible (i.e., currencies could be freely exchanged for any purpose and in any amount). This condition ensures unrestricted flow of capital. Third, each country would be able to undertake fully independent monetary policy in pursuit of domestic objectives, such as growth and inflation targets.

Unfortunately, these three conditions are not consistent. If the first two conditions were satisfied—credibly fixed exchange rates and full convertibility—then there would really be only one currency in the world. Converting from one national currency to another would have no more significance (indeed less) than deciding whether to carry coins or paper currency in your wallet. Any attempt to influence interest rates, asset prices, or inflation by adjusting the supply of one currency versus another would be futile. Thus, it should be clear that independent monetary policy is not possible if exchange rates are credibly fixed and currencies are fully convertible. *There can be no ideal currency regime.*

The impact of the currency regime on a country's ability to exercise independent monetary policy is a recurring theme in open-economy macroeconomics. It will be covered in more detail in other chapters; however, it is worthwhile to emphasize the basic point by considering what would happen in an idealized world of perfect capital mobility. If the exchange rate were credibly fixed, then any attempt to decrease default-free interest rates in

one country below those in another—that is, to undertake independent, expansionary monetary policy—would result in a potentially unlimited outflow of capital because funds would seek the higher return. The central bank would be forced to sell foreign currency and buy domestic currency to maintain the fixed exchange rate. The loss of reserves and reduction in the domestic money supply would put upward pressure on domestic interest rates until rates were forced back to equality, negating the initial expansionary policy. Similarly, contractionary monetary policy (higher interest rates) would be thwarted by an inflow of capital.

The situation is quite different, however, with a floating exchange rate. A decrease in the domestic interest rate would make the domestic currency less attractive. The resulting depreciation of the domestic currency would shift demand toward domestically produced goods (i.e., exports rise and imports fall), reinforcing the expansionary impact of the initial decline in the interest rate. Similarly, a contractionary increase in the interest rate would be reinforced by appreciation of the domestic currency.

In practice, of course, capital is not perfectly mobile and the impact on monetary policy is not so stark. The fact remains, however, that fixed exchange rates limit the scope for independent monetary policy, and national monetary policy regains potency and independence, at least to some degree, if the exchange rate is allowed to fluctuate and/or restrictions are placed on convertibility. In general, the more freely the exchange rate is allowed to float and the more tightly convertibility is controlled, the more effective the central bank can be in addressing domestic macroeconomic objectives. The downside, of course, is the potential distortion of economic activity caused by exchange rate risk and inefficient allocation of financial capital.

4.2. Historical Perspective on Currency Regimes

How currencies exchange for one another has evolved over the centuries. At any point in time, different exchange rate systems may coexist; still, there tends to be one dominant system in the world economy. Throughout most of the nineteenth century and during the early twentieth century until the start of World War I, the U.S. dollar and the UK pound sterling operated on the classical gold standard. The price of each currency was fixed in terms of gold. Gold was the numeraire[10] for each currency; therefore, it was indirectly the numeraire for all other prices in the economy. Many countries (e.g., the colonies of the United Kingdom) fixed their currencies relative to sterling and were therefore implicitly also operating on the classical gold standard.

The classical gold standard operated by what is called the *price-specie-flow mechanism*. This mechanism operated through the impact of trade imbalances on capital flows, namely gold. As countries experienced a trade surplus, they accumulated gold as payment, their domestic money supply expanded by the amount dictated by the fixed parity, prices rose, and exports fell. Similarly, when a country ran a trade deficit, there was an automatic outflow of gold, a contraction of the domestic money supply, and a fall in prices leading to increased exports.

In this system, national currencies were backed by gold. A country could print only as much money as its gold reserve warranted. The system was limited by the amount of gold, but it was self-adjusting and inspired confidence. With a fixed stock of gold, the price-specie-flow mechanism would work well. Still, new gold discoveries as well as more efficient methods of refining gold would enable a country to increase its gold reserves and increase its money supply

[10]Economists refer to the unit of account in terms of which other goods, services, and assets are priced as the *numeraire*. Under the classical gold standard, the official value of each currency was expressed in ounces of gold.

apart from the effect of trade flows. In general, however, trade flows drove changes in national money supplies.[11]

There is much disagreement among economic historians about the effect of the classical gold standard on overall macroeconomic stability. Was it destabilizing? On the one hand, monetary policy was tied to trade flows, so a country could not engage in expansionary policies when there was a downturn in the nontraded sector. On the other hand, it has been argued that tying monetary policy to trade flows kept inflation in check.

During the 1930s, the use of gold as a clearing device for settlement of trade imbalances, combined with increasing protectionism on the part of economies struggling with depression as well as episodes of deflation and hyperinflation, created a chaotic environment for world trade. As a consequence of these factors, world trade dropped by over 50 percent and the gold standard was abandoned.

In the later stages of World War II, a new system of fixed exchange rates with periodic realignments was devised by John Maynard Keynes and Harry Dexter White, representing the UK and U.S. treasuries, respectively. The Bretton Woods system, named after the town where it was negotiated, was adopted by 44 countries in 1944. From the end of the war until the collapse of the system in the early 1970s, the United States, Japan, and most of the industrialized countries of Europe maintained a system of fixed parities for exchange rates between currencies. When the parities were significantly and persistently out of line with the balancing of supply and demand, there would be a realignment of currencies, with some appreciating in value and others depreciating in value. These periodic realignments were viewed as a part of standard monetary policy.

By 1973, with chronic inflation taking hold throughout the world, most nations abandoned the Bretton Woods system in favor of a flexible exchange rate system under what are known as the Smithsonian Agreements. Milton Friedman had called for such a system as far back as the 1950s.[12] His argument was that the fixed parity system with periodic realignments would become unsustainable. When the inevitable realignments were imminent, large speculative profit opportunities would appear. Speculators would force the hand of monetary policy authorities, and their actions would distort the data needed to ascertain appropriate trade-related parities. It is better, he argued, to let the market, rather than central bank governors and treasury ministers, determine the exchange rate.

After 1973, most of the industrialized world changed to a system of flexible exchange rates. The original thinking was that the forces that caused exchange rate chaos in the 1930s—poor domestic monetary policy and trade barriers—would not be present in a flexible exchange rate regime, and therefore exchange rates would move in response to the exchange of goods and services among countries. As it turned out, however, exchange rates moved around much more than anyone expected. Academic economists and financial analysts alike soon realized that the high degree of exchange rate volatility was the manifestation of a highly liquid, forward-looking asset market.[13] Investment-driven FX transactions—for both

[11]The European inflation of the seventeenth century was an important exception. Discoveries of gold in South America led to an increase in the world gold stock and in prices throughout Europe. The impact was especially pronounced in Spain, the primary importing country. Historians have attributed the decline of the Spanish Empire in part to the loss of control of domestic prices.

[12]Friedman (1953).

[13]Whether or not FX markets satisfy recognized definitions of market efficiency—correctly reflecting all available information—is debatable (e.g., some point to evidence of trending as a clear violation of efficiency). However, there is no doubt that FX market participants attempt to incorporate new information, which is often lumpy and difficult to decipher, into their expectations about the future. Changing expectations—accurate or otherwise—affect the value that investors place on holding different currencies and, in a highly liquid market, lead to rapid and sometimes violent exchange rate movements.

long-term investment and short-term speculation—mattered much more in setting the spot exchange rate than anyone had previously imagined.

There are costs, of course, to a high degree of exchange rate volatility. These include difficulty planning without hedging exchange rate risks—a form of insurance cost, domestic price fluctuations, uncertain costs of raw materials, and short-term interruptions in financing transactions. For these reasons, in 1979 the European Community, forerunner of the European Union, opted for a system of limited flexibility, the European Exchange Rate Mechanism (ERM).

Initially, the system called for European currency values to fluctuate within a narrow band called "the snake." This did not last long. The end of the Cold War and the reunification of Germany created conditions ripe for speculative attack. In the early 1990s, the United Kingdom was in a recession and the Bank of England's monetary policy leaned toward low interest rates to stimulate economic recovery. Germany was issuing large amounts of debt to pay for reunification, and the German central bank (the Deutsche Bundesbank) opted for high interest rates to ensure price stability. Capital began to flow from sterling to deutsche marks to obtain the higher interest rate. The Bank of England tried to lean against these flows and maintain the exchange rate within the Exchange Rate Mechanism, but eventually it began to run out of marks to sell. Because it was almost certain that devaluation would be required, holders of sterling rushed to purchase marks at the old rate, and the speculative attack forced the United Kingdom out of the ERM in September 1992, only two years after it had finally joined the system.

Despite these difficulties, 1999 saw the creation of a common currency for most Western European countries, without Switzerland or the United Kingdom, called the euro.[14] The hope was that the common currency would increase transparency of prices across borders in Europe, enhance market competition, and facilitate more efficient allocation of resources. The drawback, of course, is that each member country lost the ability to manage its exchange rate and therefore to engage in independent monetary policy.

4.3. A Taxonomy of Currency Regimes

Although the pros and cons of fixed and flexible exchange rate regimes continue to be debated, many countries adopt regimes that lie somewhere between these polar cases. Countries adopt specific regimes for a variety of reasons. In some cases, the driving force is the lack of credibility with respect to sound monetary policy. A country with a history of hyperinflation may be forced to adopt a form of fixed-rate regime because its promise to maintain a sound currency with a floating rate regime would not be credible. This has been a persistent issue in Latin America. In other cases, the driving force is as much political as economic. The decision to create the euro was strongly influenced by the desire to enhance political union within the European Community, whose members had been at war with each other twice in the twentieth century.

As of April 2008, the International Monetary Fund (IMF) classified the actual (as opposed to officially stated) exchange rate regimes of its members into the eight categories shown in Exhibit 9-8.

It should be noted that global financial markets are too complex and diverse to be fully captured by this (or any other) classification system. A government's control over the domestic currency's exchange rate will depend on many factors—for example, the degree of capital controls used to prevent the free flow of funds in and out of the country. Also, even those

[14]The number of European countries adopting the euro has continued to expand since its inception; the most recent country to join the euro (as of this writing) was Estonia, on 1 January 2011.

EXHIBIT 9-8 Exchange Rate Regimes for Selected Markets[a] as of 30 April 2008

Type of Regime	Currency Anchor		
	USD	EUR	Basket/None
No separate legal tender			
Dollarized	Ecuador, El Salvador, Panama	Montenegro, San Marino	
Monetary union		EMU: Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, Netherlands, Portugal, Slovenia, Spain	
Currency board	Antigua, Hong Kong	Bosnia and Herzegovina, Bulgaria	
Fixed parity	Argentina, Belarus, Lebanon, Saudi Arabia, Venezuela, Vietnam	Croatia, Denmark	Kuwait, Libya, Russia
Target zone		Slovak Republic	Syria
Crawling peg	Bolivia, China, Iraq		Iran
Crawling band	Costa Rica		Azerbaijan
Managed float	Cambodia, Liberia, Ukraine		Algeria, Colombia, Egypt, India, Indonesia, Malaysia, Peru, Singapore, Thailand
Independent float			Australia, Canada, Chile, Hungary, Iceland, Israel, Japan, Mexico, New Zealand, Norway, Philippines, Poland, South Africa, South Korea, Sweden, Switzerland, Turkey, United Kingdom, United States

[a]The classifications are described in International Monetary Fund (2006). In some cases, the labels used by the IMF do not clearly distinguish among the regimes. Hence, the names applied here to the regimes differ somewhat from the IMF's original taxonomy.

countries classified as using an independent float regime will occasionally intervene in foreign exchange markets in order to influence the value of their domestic currency. Additionally, the specifics of exchange rate policy implementation are subject to change. (For example, the Chinese yuan was officially pegged against a basket of currencies prior to the 2008 crisis in the global financial market, but the Chinese government switched back toward focusing on the rate against the U.S. dollar after market volatility rose.)

This means that the classifications in Exhibit 9-8 are somewhat arbitrary and subject to interpretation, as well as change, over time. The important point to be drawn from this discussion

is that the prices and flows in foreign exchange markets will, to varying degrees, reflect the legal and regulatory frameworks imposed by governments, not just pure market forces. Governments have a variety of motives and tools for attempting to manage exchange rates. The taxonomy in Exhibit 9-8 can be used to help understand the main distinctions among currency regimes and the rationales for adopting them, but the specific definitions should not be interpreted too rigidly. Instead, the focus should be on the diversity of foreign exchange markets globally as well as the implications of these various currency regimes for market pricing.

4.3.1. Arrangements with No Separate Legal Tender

The IMF identifies two types of arrangements in which a country does not have its own legal tender. In the first, known as *dollarization*, the country uses the currency of another nation as its medium of exchange and unit of account. In the second, the country participates in a monetary union whose members share the same legal tender. In either case, the country gives up the ability to conduct its own monetary policy.

In principle, a country could adopt any currency as its medium of exchange and unit of account, but the main reserve currency, the U.S. dollar, is an obvious choice—hence the name *dollarization*. Many countries are dollarized: East Timor, El Salvador, Ecuador, and Panama, for example. By adopting another country's currency as legal tender, a dollarized country inherits that country's currency credibility, but not its creditworthiness. For example, although local banks may borrow, lend, and accept deposits in U.S. dollars, they are not members of the U.S. Federal Reserve System, nor are they backed by deposit insurance from the Federal Deposit Insurance Corporation. Thus, interest rates on U.S. dollars in a dollarized economy need not be, and generally are not, the same as on dollar deposits in the United States.

Dollarization imposes fiscal discipline by eliminating the possibility that the central bank will be induced to monetize government debt (i.e., to persistently purchase government debt with newly created local currency). For countries with a history of fiscal excess or lack of monetary discipline, dollarizing the economy can facilitate growth of international trade and capital flows if it creates an expectation of economic and financial stability. In the process, however, it removes another potential source of stabilization—domestic monetary policy.

The European Economic and Monetary Union (EMU) is the most prominent example of the second type of arrangement lacking separate legal tender. Each EMU member uses the euro as its currency. Although member countries cannot have their own monetary policies, they jointly determine monetary policy through their representation at the European Central Bank (ECB). As with dollarization, a monetary union confers currency credibility on members with a history of fiscal excess or a lack of monetary discipline. However, as shown by the 2010–2012 EMU sovereign debt crises, monetary union alone cannot confer creditworthiness.

4.3.2. Currency Board System

The IMF defines a *currency board system* (CBS) as:

> A monetary regime based on an explicit legislative commitment to exchange domestic currency for a specified foreign currency at a fixed exchange rate, combined with restrictions on the issuing authority to ensure fulfillment of its legal obligation. This implies that domestic currency will be issued only against foreign exchange and it remains fully backed by foreign assets.[15]

[15]International Monetary Fund (2006).

Hong Kong is the leading example of a long-standing (since 1983) currency board. U.S. dollar reserves are held to cover, at the fixed parity, the entire *monetary base*—essentially bank reserves plus all HKD notes and coins in circulation.[16] Note that HKD-denominated bank deposits are not fully collateralized by U.S. dollar reserves; to do so would mean that banks could not lend against their deposits. The Hong Kong Monetary Authority (HKMA) does not function as a traditional central bank under this system because the obligation to maintain 100 percent foreign currency reserves against the monetary base prevents it from acting as a lender of last resort for troubled financial institutions. However, it can provide short-term liquidity by lending against foreign currency collateral.

A CBS works much like the classical gold standard in that expansion and contraction of the monetary base are directly linked to trade and capital flows. As with the gold standard, a CBS works best if domestic prices and wages are very flexible, nontraded sectors of the domestic economy are relatively small, and the global supply of the reserve asset grows at a slow, steady rate consistent with long-run real growth with stable prices. The first two of these conditions are satisfied in Hong Kong. Until and unless Hong Kong selects a new reserve asset, however, the third condition depends on U.S. monetary policy.

In practice, the HKD exhibits modest fluctuations around the official parity of HKD/USD = 7.80 because the HKMA buys USD at a preannounced level slightly below the parity and sells USD above the parity. Persistent flows on one side of this convertibility zone or the other result in interest rate adjustments rather than exchange rate adjustments. Inside the zone, however, the exchange rate is determined by the market, and the HKMA is free to conduct limited monetary operations aimed at dampening transitory interest rate movements.

One of the advantages of a CBS as opposed to dollarization is that the monetary authority can earn a profit by paying little or no interest on its liability—the monetary base—and can earn a market rate on its asset—the foreign currency reserves. This profit is called *seigniorage*.[17] Under dollarization, the seigniorage goes to the country whose currency is used.

4.3.3. Fixed Parity

A simple fixed-rate system differs from a CBS in two important respects. First, there is no legislative commitment to maintaining the specified parity. Thus, market participants know that the country may choose to adjust or abandon the parity rather than endure other, potentially more painful, adjustments. Second, the target level of foreign exchange reserves is discretionary; it bears no particular relationship to domestic monetary aggregates. Thus, although monetary independence is ultimately limited as long as the exchange peg is maintained, the central bank can carry out traditional functions, such as serving as lender of last resort.

In the conventional fixed-rate system, the exchange rate may be pegged to a single currency—for example, the U.S. dollar—or to a basket index of the currencies of major trading partners. There is a band of up to ±1 percent around the parity level within which private flows are allowed to determine the exchange rate. The monetary authority stands ready to spend its foreign currency reserves, or buy foreign currency, in order to maintain the rate within these bands.

---

[16]For a description of Hong Kong's currency board system, see Hong Kong Monetary Authority (2005).

[17]More generally, seigniorage is the profit earned when the value of money issued exceeds the cost of producing it. For physical currency, seigniorage arises when a coin is minted for a fraction of its face value and then issued (sold) at its face value.

The credibility of the fixed parity depends on the country's willingness and ability to offset imbalances in private-sector demand for its currency. Both excess and deficient private demand for the currency can exert pressure to adjust or abandon the parity. Excess private demand for the domestic currency implies a rapidly growing stock of foreign exchange reserves, expansion of the domestic money supply, and potentially accelerating inflation. Deficient demand for the currency depletes foreign exchange reserves and exerts deflationary pressure on the economy. If market participants believe the foreign exchange reserves are insufficient to sustain the parity, then that belief may be self-fulfilling because the resulting speculative attack will drain reserves and may force an immediate devaluation. Thus, the level of reserves required to maintain credibility is a key issue for a simple fixed exchange rate regime.

4.3.4. Target Zone

A target zone regime has a fixed parity with fixed horizontal intervention bands that are somewhat wider—up to ± 2 percent around the parity—than in the simple fixed parity regime. The wider bands provide the monetary authority with greater scope for discretionary policy.

4.3.5. Active and Passive Crawling Pegs

Crawling pegs for the exchange rate—usually against a single currency, such as the U.S. dollar—were common in the 1980s in Latin America, particularly in Brazil, during the high inflation periods. To prevent a run on the U.S. dollar reserves, the exchange rate was adjusted frequently (weekly or daily) to keep pace with the inflation rate. Such a system was called a passive crawl. An adaptation used in Argentina, Chile, and Uruguay was the active crawl: The exchange rate was preannounced for the coming weeks, with changes taking place in small steps. The aim of the active crawl was to manipulate expectations of inflation. Because the domestic prices of many goods were directly tied to import prices, announced changes in the exchange rate would effectively signal future changes in the inflation rate of these goods.

4.3.6. Fixed Parity with Crawling Bands

A country can also have a fixed central parity with crawling bands. Initially, a country may fix its rates to a foreign currency to anchor expectations about future inflation but then gradually permit more and more flexibility in the form of a preannounced widening band around the central parity. Such a system has the desirable property of allowing a gradual exit strategy from the fixed parity. A country might want to introduce greater flexibility and greater scope for monetary policy, but it may not yet have the credibility or financial infrastructure for full flexibility. So it maintains a fixed parity with slowly widening bands.

4.3.7. Managed Float

A country may simply follow an exchange rate policy based on either internal or external policy targets—intervening or not to achieve trade balance, price stability, or employment objectives. Such a policy, often called dirty floating, invites trading partners to respond likewise with their exchange rate policies and potentially decreases stability in foreign exchange markets as a whole. The exchange rate target, in terms of either a level or a rate of change, is typically not explicit.

4.3.8. Independently Floating Rates

In the case of independently floating rates, the exchange rate is left to market determination and the monetary authority is able to exercise independent monetary policy aimed at achieving

such objectives as price stability and full employment. The central bank also has latitude to act as a lender of last resort to troubled financial institutions, if necessary.

It should be clear from recent experience that the concepts of float, managed float, crawl, and target zone are not hard-and-fast rules. Central banks do occasionally engage, implicitly or explicitly, in regime switches—even in countries nominally following an independently floating exchange rate regime. For example, when the U.S. dollar appreciated in the mid-1980s with record U.S. trade deficits, then U.S. Treasury Secretary James Baker engineered the Plaza Accord, in which Japan and Germany agreed to an appreciation of their currencies against the U.S. dollar. (The Plaza Accord is so named because it was negotiated at the Plaza Hotel in New York City.) This 1985 policy agreement involved a combination of fiscal and monetary policy measures by the countries involved as well as direct intervention in foreign exchange markets. The Plaza Accord was a clear departure from a pure independently floating exchange rate system.

There are more recent examples of government intervention in foreign exchange markets. In September 2000, the European Central Bank, the Federal Reserve Board, the Bank of Japan, the Bank of England, and the Bank of Canada engaged in a concerted intervention in order to support the value of the euro, a freely floating currency that was then under pressure within foreign exchange markets. (This intervention was described as "concerted" because it was prearranged and coordinated among the central banks involved.) During 2010, many countries engaged in unilateral intervention to prevent the rapid appreciation of their currencies against the U.S. dollar. Several of these countries also employed various fiscal and regulatory measures (for example, taxes on capital inflows) in order to further affect exchange rate movements.

The important point to draw from this discussion is that exchange rates not only reflect private-sector market forces but will also, to varying degrees, be influenced by the legal and regulatory framework (currency regimes) within which foreign exchange markets operate. Moreover, they will occasionally be influenced by government polices (fiscal, monetary, and intervention) intended to manage exchange rates. All of these can vary widely among countries and are subject to change with time.

Nonetheless, the most widely traded currencies in foreign exchange markets (the U.S. dollar, yen, euro, UK pound, Swiss franc, and the Canadian and Australian dollars) are typically considered to be free floating, although subject to relatively infrequent intervention.

EXAMPLE 9-7 Currency Regimes

An investment adviser in Los Angeles, California, is meeting with a client who wishes to diversify her portfolio by including more international investments. In order to evaluate the suitability of international diversification for the client, the adviser attempts to explain some of the characteristics of foreign exchange markets. The adviser points out that countries often follow different exchange rate regimes, and the choice of regime will affect the performance of their domestic economy, as well as the amount of foreign exchange risk posed by international investments.

The client and her adviser discuss potential investments in Hong Kong, Panama, and Canada. The adviser notes that the currency regimes of these countries are a

currency board, dollarization, and a free float, respectively. The adviser tells his client that these regimes imply different degrees of foreign exchange risk for her portfolio.

The discussion between the investment adviser and his client then turns to potential investments in other countries with different currency regimes. The adviser notes that some countries follow fixed parity regimes against the U.S. dollar. The client asks whether a fixed parity regime would imply less foreign currency risk for her portfolio than would a currency board. The adviser replies: "Yes, a fixed parity regime means a constant exchange rate and is more credible than a currency board."

The adviser goes on to explain that some countries allow their exchange rates to vary, although with different degrees of foreign exchange market intervention to limit exchange rate volatility. Citing examples, he notes that China has a crawling peg regime with reference to the U.S. dollar, but the average daily percentage changes in the China/U.S. exchange rate are very small compared with the average daily volatility for a freely floating currency. The adviser also indicates that Denmark has a target zone regime with reference to the euro, and South Korea usually follows a freely floating currency regime but sometimes switches to a managed float regime. The currencies of China, Denmark, and South Korea are the yuan renminbi (CNY), krone (DKK), and won (KRW), respectively.

1. Based solely on the exchange rate risk the client would face, what is the correct ranking (from most to least risky) of the following investment locations?
   A. Panama, Canada, Hong Kong
   B. Canada, Hong Kong, Panama
   C. Hong Kong, Panama, Canada
2. Based solely on their foreign exchange regimes, which country is *least likely* to import inflation or deflation from the United States?
   A. Canada
   B. Panama
   C. Hong Kong
3. The adviser's statement about fixed parity regimes is incorrect with regard to:
   A. credibility.
   B. a constant exchange rate.
   C. both a constant exchange rate and credibility.
4. Based on the adviser's categorization of China's currency regime, if the USD is depreciating against the KRW, then it is *most likely* correct that the CNY is:
   A. fixed against the KRW.
   B. appreciating against the KRW.
   C. depreciating against the KRW.
5. Based on the adviser's categorization of Denmark's currency regime, it would be *most* correct to infer that:
   A. the krone is allowed to float against the euro within fixed bands.
   B. the Danish central bank will intervene if the exchange rate strays from its target level.
   C. the target zone will be adjusted periodically in order to manage inflation expectations.

6. Based on the adviser's categorization of South Korea's currency policy, it would be *most* correct to infer that:
   A. the South Korean central bank is engineering a gradual exit from a fixed-rate regime.
   B. the government is attempting to peg the exchange rate within a predefined zone.
   C. the won is allowed to float, but with occasional intervention by the South Korean central bank.

Solution to 1: B is correct. The CAD/USD exchange rate is a floating exchange rate, and Canadian investments would therefore carry exchange rate risk for a U.S.-based investor. Although Hong Kong follows a currency board system, the HKD/USD exchange rate nonetheless does display some variation, albeit much less than in a floating exchange rate regime. In contrast, Panama has a dollarized economy (i.e., it uses the U.S. dollar as the domestic currency); therefore, there is no foreign exchange risk for a U.S. investor.

Solution to 2: A is correct. The Canadian dollar floats independently against the U.S. dollar, leaving the Bank of Canada able to adjust monetary policy to maintain price stability. Neither Hong Kong (currency board) nor Panama (dollarized) can exercise independent monetary policy to buffer its economy from the inflationary/deflationary consequences of U.S. monetary policy.

Solution to 3: C is correct. A fixed parity regime does not mean that the exchange rate is rigidly fixed at a constant level. In practice, both a fixed parity regime and a currency board allow the exchange rate to vary within a band around the country's stated parity level. Thus, both regimes involve at least a modest amount of exchange rate risk. The fixed parity regime exposes the investor to the additional risk that the country may be unable or unwilling to maintain the parity. In a fixed parity regime, the level of foreign currency reserves is discretionary and typically only a small fraction of the domestic money supply. With no legal obligation to maintain the parity, the country may adjust the parity (devalue or revalue its currency) or allow its currency to float if doing so is deemed to be less painful than other adjustment mechanisms (e.g., fiscal restraint). In contrast, a currency board entails a legal commitment to maintain the parity and to fully back the domestic currency with reserve currency assets. Hence, there is little risk that the parity will be abandoned.

Solution to 4: C is correct. If the CNY is subject to a crawling peg with very small daily adjustments versus the USD and the USD is depreciating against the KRW, then the CNY would most likely be depreciating against the KRW as well. In fact, this was an important issue in foreign exchange markets through the latter part of 2010: As the USD depreciated against most Asian currencies (and less so against the CNY), many Asian countries felt that they were losing their competitive export advantage because the CNY was so closely tied to the USD. This led many Asian countries to intervene in FX markets against the strength of their domestic currencies in order not to lose an export pricing advantage against China.

Solution to 5: A is correct. A target zone means that the exchange rate between the euro and Danish krone (DKK) will be allowed to vary within a fixed band (as of 2010, the target zone for the DKK/EUR is a ±2.5 percent band). This does not mean that the DKK/EUR rate is fixed at a certain level (B is incorrect) or that the target zone will vary in order to manage inflation expectations (this is a description of a crawling peg, which makes C incorrect).

Solution to 6: C is correct. Similar to the monetary authorities responsible for many of the world's major currencies, the South Korean policy typically involves letting market forces determine the exchange rate (an independent floating rate regime). But this approach does not mean that market forces are the sole determinant of the won exchange rate. As with most governments, the South Korean policy is to intervene in foreign exchange markets when movements in the exchange rate are viewed as undesirable (a managed float). For example, during the latter part of 2010, South Korea and many other countries intervened in foreign exchange markets to moderate the appreciation of their currencies against the U.S. dollar. Answer A describes a fixed parity with a crawling bands regime, and B describes a target zone regime; both answers are incorrect.

5. EXCHANGE RATES, INTERNATIONAL TRADE, AND CAPITAL FLOWS

Just as a family that spends more than it earns must borrow or sell assets to finance the excess, a country that imports more goods and services than it exports must borrow from foreigners or sell assets to foreigners to finance the trade deficit. Conversely, a country that exports more goods and services than it imports must invest the excess either by lending to foreigners or by buying assets from foreigners. Thus, a trade deficit must be exactly matched by an offsetting *capital account* surplus, and a trade surplus must be matched by a capital account deficit.[18] This implies that any factor that affects the trade balance must have an equal and opposite impact on the capital account, and vice versa. To put this somewhat differently, the *impact of exchange rates and other factors on the trade balance must be mirrored by their impact on capital flows*; they cannot affect one without affecting the other.

Using a fundamental identity from macroeconomics, the relationship between the trade balance and expenditure/saving decisions can be expressed as:[19]

$$X - M = (S - I) + (T - G)$$

[18]In official balance of payments accounts, investment/financing flows are separated into two categories: the capital account and the financial account. Because the technical distinction is immaterial for present purposes, we will simply refer to the balance of investment/financing flows as the capital account. Similarly, we ignore the technical distinction between the trade balance and the current account balance. The details of balance of payments accounting were presented in Chapter 8, "International Trade and Capital Flows."

[19]This relationship was developed in Chapter 5, "Aggregate Output, Prices, and Economic Growth."

where X represents exports, M is imports, S is private savings, I is investment in plant and equipment, T is taxes net of transfers, and G is government expenditure. From this relationship, we can see that a trade surplus ($X > M$) must be reflected in a fiscal surplus ($T > G$), an excess of private saving over investment ($S > I$), or both. Because a fiscal surplus can be viewed as government saving, we can summarize this relationship more simply by saying that a trade surplus means the country saves more than enough to fund its investment (I) in plant and equipment. The excess saving is used to accumulate financial claims on the rest of the world. Conversely, a trade deficit means the country does not save enough to fund its investment spending (I) and must reduce its net financial claims on the rest of the world.

Although this identity provides a key link between real expenditure/saving decisions and the aggregate flow of financial assets into or out of a country, it does not tell us what type of financial assets will be exchanged or in what currency they will be denominated. All that can be said is that asset prices and exchange rates at home and abroad must adjust so that all financial assets are willingly held by investors.

If investors anticipate a significant change in an exchange rate, they will try to sell the currency that is expected to depreciate and buy the currency that is expected to appreciate. This implies an incipient (i.e., potential) flow of capital from one country to the other, which must either be accompanied by a simultaneous shift in the trade balance or be discouraged by changes in asset prices and exchange rates. Because expenditure/saving decisions and prices of goods change much more slowly than financial investment decisions and asset prices, most of the adjustment usually occurs within the financial markets. That is, *asset prices and exchange rates adjust so that the potential flow of financial capital is mitigated and actual capital flows remain consistent with trade flows*. In a fixed exchange rate regime, the central bank offsets the private capital flows in the process of maintaining the exchange rate peg, and the adjustment occurs in other asset prices, typically interest rates, until and unless the central bank is forced to allow the exchange rate to adjust.[20] In a floating exchange rate regime, the main adjustment is often a rapid change in the exchange rate that dampens an investor's conviction that further movement will be forthcoming. Thus, *capital flows—potential and actual—are the primary determinant of exchange rate movements in the short to intermediate term*. Trade flows become increasingly important in the longer term as expenditure/saving decisions and the prices of goods and services adjust.

With the correspondence between the trade balance and capital flows firmly established, we can now examine the impact of exchange rate changes on the trade balance from two perspectives. The first approach focuses on the effect of changing the relative prices of domestic and foreign goods. This approach, which is called the *elasticities approach*, highlights changes in the composition of spending. The second approach, called the *absorption approach*, focuses on the impact of exchange rates on aggregate expenditure/saving decisions.

5.1. Exchange Rates and the Trade Balance: The Elasticities Approach

The effectiveness of devaluation (in a fixed system) or depreciation (in a flexible system) of the currency for reducing a trade deficit depends on well-behaved demand and supply curves for goods and services. The condition that guarantees that devaluations improve the trade balance is called the Marshall–Lerner condition. The usual statement of this condition assumes that

[20]A classic example of this occurred in September 1992, when the United Kingdom was forced to withdraw from the European Exchange Rate Mechanism, the forerunner of the current European Economic and Monetary Union (EMU).

trade is initially balanced. We will present a generalization of the condition that allows for an initial trade imbalance and hence is more useful in addressing whether exchange rate movements will correct such imbalances.

Recall from microeconomics that the price elasticity of demand is given by:[21]

$$\varepsilon = -\frac{\%\ \textit{change in quantity}}{\%\ \textit{change in price}} = -\frac{\%\Delta Q}{\%\Delta P}$$

For example, a demand elasticity of 0.6 means that quantity demanded increases by 6 percent if price declines by 10 percent. Note that the elasticity of demand is defined so that it is a positive number. Because expenditure (R) equals price multiplied by quantity ($P \times Q$), by rearranging the preceding expression to solve and substitute for $\%\Delta Q$, we can see that:

$$\%\ \textit{change in expenditure} = \%\Delta R = \%\Delta P + \%\Delta Q = (1 - \varepsilon)\%\Delta P$$

From this we can see that an increase in price decreases expenditure if $\varepsilon > 1$, but it increases expenditure if $\varepsilon < 1$. By convention, if $\varepsilon > 1$, demand is described as being "elastic," whereas if $\varepsilon < 1$, demand is described as "inelastic."

The basic idea behind the Marshall–Lerner condition is that demand for imports and exports must be sufficiently price sensitive that increasing the relative price of imports increases the difference between export receipts and import expenditures. The generalized Marshall–Lerner condition is:

$$\omega_X \varepsilon_X + \omega_M(\varepsilon_M - 1) > 0$$

where ω_X and ω_M are the shares of exports and imports, respectively, in total trade (i.e., imports + exports) and ε_X and ε_M are the price elasticities of foreign demand for domestic country exports and domestic country demand for imports, respectively. Note that $(\omega_X + \omega_M) = 1$ and that an initial trade deficit implies $\omega_M > \omega_X$. If this condition is satisfied, a devaluation/depreciation of the domestic currency will move the trade balance toward surplus.

The first term in the generalized Marshall–Lerner condition reflects the change in export receipts assuming the domestic currency price of exports is unchanged (i.e., foreigners are billed in the domestic currency). It will be positive as long as export demand is not totally insensitive to price. Depreciation of the domestic currency makes exports cheaper in foreign currency and induces an increase in the quantity demanded by foreigners. This is reflected by the elasticity ε_X. There is no direct price impact on domestic currency export revenue because the domestic currency price is assumed to be unchanged. Hence, the percentage change in export revenue corresponding to a 1 percent depreciation of the currency is simply ε_X. The second term in the generalized Marshall–Lerner condition reflects the impact on import expenditures. Assuming that imports are billed in a foreign currency, the domestic currency price of imports rises as the domestic currency depreciates. The direct price effect increases import expenditures, while the induced reduction in the quantity of imports decreases import expenditures. The net effect depends on the elasticity of import demand, ε_M. Import expenditure declines only if import demand is elastic (i.e., $\varepsilon_M > 1$).

[21] See Chapter 1, "Demand and Supply Analysis: Introduction."

Examination of the generalized Marshall–Lerner condition indicates that more elastic demand—for either imports or exports—makes it more likely that the trade balance will improve. Indeed, if the demand for imports is elastic, $\varepsilon_M > 1$, then the trade balance will definitely improve. It should also be clear that the elasticity of import demand becomes increasingly important, and the export elasticity less important, as the initial trade deficit gets larger—that is, as ω_M increases. In the special case of initially balanced trade, $\omega_X = \omega_M$, the condition reduces to $(\varepsilon_X + \varepsilon_M > 1)$, which is the classic Marshall–Lerner condition.

Exhibit 9-9 illustrates the impact of depreciation on the trade balance. For ease of reference, we assume the domestic currency is the euro. A 10 percent depreciation of the euro makes imports 10 percent more expensive in euro terms. With an import elasticity of 0.65, this induces a 6.5 percent reduction in the quantity of imports. But import expenditures increase by 3.5 percent [10% × (1 – 0.65)] or €21,000,000 because the drop in quantity is not sufficient to offset the increase in price. On the export side, the euro price of exports does not change, but the foreign currency price of exports declines by 10 percent. This induces a 7.5 percent increase in the quantity of exports, given an elasticity of 0.75. The euro value of exports therefore increases by 7.5 percent or €30,000,000. The net effect is a €9,000,000 improvement in the trade balance and a €51,000,000 increase in total trade.

The balance of trade improves after the depreciation of the euro because the Marshall–Lerner condition is satisfied: The increase in the euro-value of exports exceeds the increase in the value of imports. Based on the data in Exhibit 9-9, $\omega_M = 0.6$ (i.e., 600,000,000/1,000,000,000) and $\omega_X = 0.4$ (i.e., 1 – 0.6). Thus, the Marshall–Lerner equation is greater than zero:

$$\omega_X \varepsilon_X + \omega_M(\varepsilon_M - 1) = 0.4 \times 0.75 + 0.6(0.65 - 4) = 0.09$$

The elasticity of demand for any good or service depends on at least four factors: (1) the existence or absence of close substitutes, (2) the structure of the market for that product (e.g., a monopoly or perfect competition), (3) its share in people's budgets, and (4) the nature of the product and its role in the economy. Demand for a product with close substitutes is highly

EXHIBIT 9-9 Marshall–Lerner Condition with a 10 Percent Depreciation of Domestic Currency (€)

Assumptions	Exports	Imports
Demand elasticity	0.75	0.65
Percent price change		
In domestic currency (€)	0	10%
In foreign currency	−10%	0
Results	**Initial Value (€)**	**Change (€)**
Exports	400,000,000	30,000,000
Imports	600,000,000	21,000,000
Trade balance	−200,000,000	9,000,000
Total trade	1,000,000,000	51,000,000

price sensitive, whereas demand for a unique product tends to be much less elastic. The demand curve faced by any producer also depends on the nature and level of competition among producers of that product. If there are many sellers of identical products, then each producer faces highly elastic demand for its output even if global demand for that product is insensitive to price. A producer that is able to differentiate its product, perhaps through branding, faces somewhat less elastic demand. In markets with only a few sellers, each producer faces demand that is highly dependent upon strategic maneuvers by its competitors. If competitors match price decreases but not increases, then the producer loses market share by raising its price but fails to gain market share by reducing its price.

Price changes have two effects on demand. The *substitution effect* refers to changes in the composition of spending across different products. As a product gets more expensive relative to other products, customers demand less of it, and as it gets cheaper, customers demand more of it. This is what people usually think of first when they consider the effect of a price change. The *income effect* refers to the fact that price changes affect real purchasing power. When the price of a good rises, people's purchasing power is reduced, and when the price falls, their purchasing power is increased. The strength of this effect depends on the product's share in people's budgets—the more important the product, the stronger the income effect. The income effect also depends on the nature of the product. The demand for luxuries is highly sensitive to income, whereas the demand for necessities is fairly insensitive to income.

To illustrate the differential impact of the two drivers of the income effect—share of expenditure and nature of the product—consider the demand for food. Clearly, food is a necessity. Based on this fact, we would expect demand to be inelastic. However, the share of expenditures that go to food varies across countries. In poor countries, food represents a much larger share of expenditures than in rich countries. Hence, all else being equal, we would expect the demand for food to be more price-elastic in poorer countries. Of course, even in rich countries, the composition of spending on food may change considerably even if overall demand for food does not.

A significant portion of international trade occurs in intermediate products—products that are used as inputs into the production of other goods. Demand for these products derives from supply and demand decisions for the final products. However, the same basic considerations apply for intermediate products as for final products. Are there close substitutes for an intermediate product in the production process? If not, its demand will tend to be less elastic than would be the case if there were readily available substitutes. How important is it to the overall economy? All else being equal, the larger its share in overall production costs for the economy, the bigger its impact on production decisions and therefore the more price-elastic its derived demand. Oil is a classic example of a widely used input with few readily adoptable substitutes, at least in the short run. Lack of substitutes tends to make oil demand price-inelastic. However, it is so important in modern industrial economies that changes in its price can induce expansion or contraction of aggregate output. This makes short-run oil demand somewhat more elastic—at least for significant price changes. In the longer run, the feasibility of substitution among energy sources enhances the price sensitivity of oil demand.

Exhibit 9-10 shows estimates of demand elasticity for various products. The estimates range from essentially zero for pediatric doctor visits—a necessity for which there is virtually no substitute—to 3.8 for Coca-Cola, a specific brand for which there are many substitutes. Note that the elasticity of demand for soft drinks in general is much lower than for Coca-Cola, roughly 0.9. The elasticity of demand for rice in Japan versus in Bangladesh clearly illustrates the impact of expenditure share on price sensitivity. Similarly, although air travel for pleasure (a luxury) is quite price-elastic, demand for first-class air travel is fairly insensitive to price.

EXHIBIT 9-10 Estimates of Demand Elasticities

Product Description	Elasticity	Rationale/Comment
Travel and transport		
Airline travel (U.S.)		
For pleasure	1.5	Luxury
First class	0.3	Business and wealthy travelers
Car fuel (U.S., long term)	0.6	
Bus travel (U.S.)	0.2	
Ford compact car	2.8	Large purchase; specific brand
Food and beverages		
Rice		Necessity; staple food
Bangladesh	0.8	Poor country
Japan	0.3	Wealthy country
Soft drinks		
All	0.8–1.0	
Coca-Cola	3.8	Specific brand; competitive market
Medical care (U.S.)		
Health insurance	0.3	
Pediatric doctor visit	0.0–0.1	No good substitute
Materials and energy		Necessary inputs
Steel	0.2–0.3	
Oil	0.4	

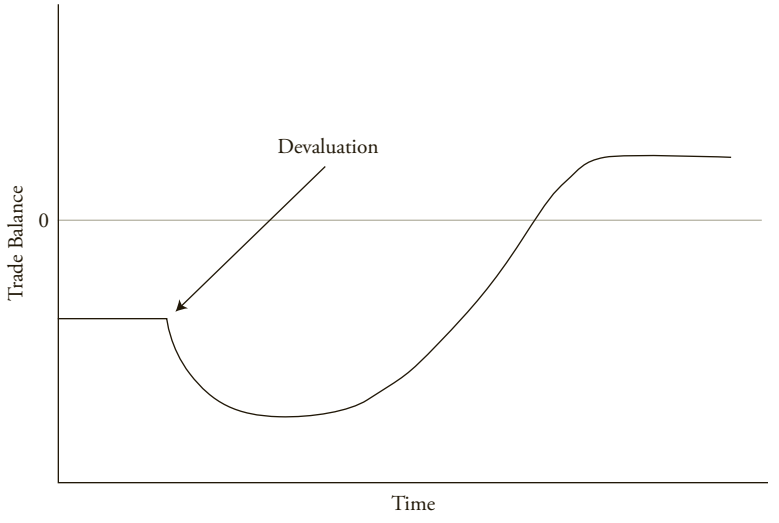
Sources: Various studies cited in Wikipedia, "Price Elasticity of Demand," as of December 2010 (http://en.wikipedia.org/wiki/Price_elasticity_of_demand).

This is most likely because many first-class passengers are either traveling on business (presumably deemed to have high value added) or wealthy enough that the cost of first-class airfare is inconsequential.

In practice, most countries import and export a variety of products. Hence, the overall price elasticities of their imports and exports reflect a composite of the products they trade. In conjunction with the Marshall–Lerner condition, our review of the factors that determine price elasticities suggests that exchange rate changes will be a more effective mechanism for trade balance adjustment if a country imports and exports the following:

- Goods for which there are good substitutes.
- Goods that trade in competitive markets.
- Luxury goods, rather than necessities.
- Goods that represent a large portion of consumer expenditures or a large portion of input costs for final producers.

EXHIBIT 9-11 Trade Balance Dynamics: The J-Curve



Note that each of these conditions is associated with higher demand elasticities (ε_X and ε_M).

Even when the Marshall–Lerner condition is satisfied, it is still possible that devaluation (in a fixed parity regime) or depreciation (in a floating regime) of the currency will initially make the trade balance worse before making it better. This effect, called the J-curve effect, is illustrated in Exhibit 9-11.

In the very short run, the J-curve reflects the order delivery lags that take place in import and export transactions. Imagine a clothing importer in New York. Orders are placed in January for French spring fashions. Market forces cause the dollar to depreciate in February, but contracts were already signed for payment in euros. When the fashions arrive in March, more dollars have to go out to pay for the order signed in euros. Thus, the trade balance gets worse. However, after the depreciation, the clothing importer has to put in new orders for summer fashions. As a result of the currency depreciation, the French summer fashions are now more expensive, so the clothing store cuts the demand for imported clothes from France. The depreciation eventually improves the trade balance, even though it initially made it worse.

A J-curve pattern may also arise if short-term price elasticities do not satisfy the Marshall–Lerner condition but long-term elasticities do satisfy it. As noted earlier in the case of oil, significant changes in spending patterns often take time. Thus, the trade balance may worsen initially and then gradually improve following a depreciation of the currency as firms and consumers adapt.

5.2. Exchange Rates and the Trade Balance: The Absorption Approach

The elasticities approach focuses on the expenditure-switching effect of changing the relative prices of imports and exports. It is essentially a microeconomic view of the relationship between exchange rates and the trade balance. The absorption approach adopts an explicitly macroeconomic view of this relationship.

Recall that the trade balance is equal to the country's saving, including the government fiscal balance, minus its investment in new plants and equipment. Equivalently, it is equal to the difference between income (gross domestic product [GDP]) and domestic expenditure, or absorption. Thus, in order to move the trade balance toward surplus, a devaluation/depreciation of the domestic currency must increase income relative to expenditure or, equivalently, increase national saving relative to investment in physical capital.

If there is excess capacity in the economy, then by switching demand toward domestically produced goods and services, depreciation of the currency can increase output/income. Because some of the additional income will be saved, income rises relative to expenditure and the trade balance improves. If the economy is at full employment, however, the trade balance cannot improve unless domestic expenditure declines. If expenditure does not decline, then the depreciation will put upward pressure on domestic prices until the stimulative effect of the exchange rate change is negated by the higher price level and the trade balance reverts to its original level.

How might depreciation of the currency reduce domestic expenditure relative to income? The main mechanism is a wealth effect. A weaker currency reduces the purchasing power of domestic-currency-denominated assets (including the present value of current and future earned income). Households respond by reducing expenditure and increasing saving in order to rebuild their wealth. Of course, as real wealth is rebuilt, the effect on saving is likely to be reversed—resulting in only a temporary improvement in the trade balance. Thus, in the absence of excess capacity in the economy, currency depreciation is likely to provide only a temporary solution for a chronic trade imbalance. Lasting correction of the imbalance requires more fundamental changes in expenditure/saving behavior (e.g., a policy shift that improves the fiscal balance or an increase in saving relative to capital investment induced by an increase in real interest rates).

The absorption approach also reminds us that currency depreciation cannot improve the trade balance unless it also induces a corresponding change in the capital account. Not only must domestic saving increase, but that saving must also be willingly channeled into buying financial assets from foreigners. All else being equal, this implies that foreign and domestic asset prices must change such that foreign assets become relatively more attractive and domestic assets relatively less attractive to both foreign and domestic investors.

EXAMPLE 9-8 Exchange Rates and the Trade Balance

An analyst at a foreign exchange dealing bank is examining the exchange rate for the Australian dollar (AUD), which is a freely floating currency. Currently, Australia is running a trade surplus with the rest of the world, primarily reflecting strong demand for Australian resource exports generated by rapid growth in emerging market economies in the western Pacific region. In turn, Australia imports food and energy from a variety of foreign countries that compete with each other as well as with Australian producers of these products. The analyst uses data in the following table to estimate the effect of changes in the AUD exchange rate on Australia's balance of trade.

	Volume (AUD billions)	Demand Elasticity
Exports	200	0.3
Imports	180	0.6

The analyst's research report on this topic notes that the mix of products that Australia imports and exports seems to be changing and that this will affect the relationship between the exchange rates and the trade surplus. The proportion of Australian exports accounted for by fine wines is increasing. These are considered a luxury good and must compete with increased wine exports from comparable-producing regions (such as Chile and New Zealand). At the same time, rising income levels in Australia are allowing the country to increase the proportion of its imports accounted for by luxury goods, and these represent a rising proportion of consumer expenditures. The analyst's report states: "Given the changing export mix, an appreciation of the currency will be more likely to reduce Australia's trade surplus. In contrast, the changing import mix will have the opposite effect."

1. Given the data in the table, an appreciation in the AUD will:
   A. cause the trade balance to increase.
   B. cause the trade balance to decrease.
   C. have no effect on the trade balance.
2. All else being equal, an appreciation in the AUD will be *more likely* to reduce the trade surplus if:
   A. the demand elasticities for imports and exports increase.
   B. the demand elasticity for exports and the export share in total trade decrease.
   C. the demand elasticity for imports decreases and the import share in total trade increases.
3. All else being equal, an appreciation in the AUD will be *more likely* to reduce the trade surplus if it leads to an increase in Australian:
   A. tax receipts.
   B. private-sector investment.
   C. government budget surpluses.
4. The report's statement about the effect of changing import and export mixes is *most likely*:
   A. correct.
   B. incorrect with respect to the import effect.
   C. incorrect with respect to the export effect.
5. Suppose the Australian government imposed capital controls that prohibited the flow of financial capital into or out of the country. What impact would this have on the Australian trade balance?
   A. The trade surplus would increase.
   B. The trade balance would go to zero.
   C. The trade balance would not necessarily be affected.
6. Suppose the Australian government imposed capital controls that prohibited the flow of financial capital into or out of the country. The impact on the trade balance, if any, would most likely take the form of:
   A. a decrease in private saving.
   B. a decrease in private investment.
   C. an increase in the government fiscal balance.

Solution to 1: A is correct. As the AUD appreciates, the price of exports to *offshore buyers* goes up and they demand fewer of them; hence, the AUD-denominated revenue from exports decreases. (Although export demand is inelastic, or $\varepsilon_X < 1$, recall that the *Australian* price of these exports is assumed not to have changed, so the amount of export revenue received by Australia, in AUD terms, unambiguously declines as the quantity of exports declines.) Australian expenditure for imports also declines. Although the price of imports declines as the AUD appreciates, the Australians do not increase their import purchases enough to lead to higher expenditures. This is because import demand is also inelastic ($\varepsilon_M < 1$). This effect on import expenditure can be seen from $\%\Delta R_M = (1 - \varepsilon_M)\ \%\Delta P_M$, where $\%\Delta P_M$ is negative (import prices are declining) and import demand is inelastic, so $(1 - \varepsilon_M) > 0$. With both import expenditures and export revenues declining, the net effect on the trade balance comes down to the relative size of the import and export weights (ω_M and ω_X, respectively). In this case, $\omega_X = 0.53$ (i.e., 200/380) and $\omega_M = 0.47$ (i.e., 180/380). Putting this into the Marshall–Lerner equation leads to:

$$\omega_X \varepsilon_X + \omega_M(\varepsilon_M - 1) = 0.53 \times 0.3 + 0.47(0.6 - 1) = -0.03$$

Because the Marshall–Lerner condition is not satisfied, exchange rate movements do not move the trade balance in the expected direction (i.e., appreciation of the currency does not decrease the trade balance, nor does depreciation of the currency increase the trade balance). However, note that with different import/export weights and the same elasticities, the Marshall–Lerner condition would be met. In particular, the condition would be met for any value of ω_X greater than 4/7 (≈ 0.571).

Solution to 2: A is correct. The basic intuition of the Marshall–Lerner condition is that in order for an exchange rate movement to rebalance trade, the demands for imports and exports must be sufficiently price sensitive (i.e., they must have sufficiently high elasticities). However, the relative share of imports and exports in total trade must also be considered. The generalized Marshall–Lerner condition requires:

$$\omega_X \varepsilon_X + \omega_M(\varepsilon_M - 1) > 0$$

An increase in both ε_X and ε_M will clearly make this expression increase (A is correct). In contrast, a decrease in both ω_X and ε_X tends to make the expression smaller (B is incorrect).* If ε_M decreases and ω_M increases, import demand will respond less to an exchange rate movement and will have a larger role in determining the trade balance (C is incorrect).

Solution to 3: B is correct. An Australian trade surplus means that Australia is spending less than it earns and is accumulating claims on foreigners. Equivalently, Australian saving, inclusive of both private saving and the government fiscal balance, is more than sufficient to fund Australian private-sector investment. The relationship between the trade balance and expenditure/saving decisions is given by:

$$X - M = (S - I) + (T - G) > 0$$

For Australia's trade balance to decline, it must save less (S down), invest more (I up), decrease its fiscal balance ($T - G$ down), or some combination of these. Increasing tax receipts (T up) increases rather than decreases the fiscal balance, so answer A is incorrect. Similarly, answer C, increasing the government budget surplus, is incorrect. Increasing private investment (I up) does decrease the trade balance, so answer B is correct.

Solution to 4: B is correct. As Australian exports become more dominated by luxury goods that face highly competitive market conditions, the elasticity of export demand (ε_X) is likely to be increasing. Increasing export elasticity makes the trade surplus more responsive to an AUD appreciation (the increase in ε_X will tend to increase the computed value for the Marshall–Lerner equation). Similarly, as Australian imports become more dominated by luxury goods that are an increasing proportion of household expenditure, import elasticity (ε_M) will most likely increase. This will also tend to increase the computed value for the Marshall–Lerner equation.

Solution to 5: B is correct. A trade deficit or surplus must be exactly matched by an offsetting capital account surplus or deficit. Anything that impacts the trade balance must impact the capital account, and vice versa. If capital flows are prohibited, then both the capital account and the trade balance must be zero.

Solution to 6: A is correct. The trade balance must go to zero. An increase in the fiscal balance implies an increase in the existing trade surplus, so answer C is incorrect. A decrease in private investment will also cause an increase in the trade surplus, so answer B is incorrect. A decrease in private saving will decrease the trade surplus as required, so answer A is correct: A decrease in saving will most likely reflect a decline in national income, especially the profit component, as export demand is choked off by the inability to extend credit to foreigners.

*Because $\omega_M = 1 - \omega_X$ and $\varepsilon_M < 1$ in this example, a decrease in ω_X also decreases the second terms, $\omega_M(\varepsilon_M - 1)$, in the Marshall–Lerner condition.

6. SUMMARY

Foreign exchange markets are crucial for understanding both the functioning of the global economy and the performance of investment portfolios. In this chapter, we have described the diverse array of FX market participants and have introduced some of the basic concepts necessary to understand the structure and functions of these markets. The reader should be able to understand how exchange rates—both spot and forward—are quoted and be able to calculate cross exchange rates and forward rates. We also have described the array of exchange rate regimes that characterize foreign exchange markets globally and how these regimes determine the flexibility of exchange rates, and hence the degree of foreign exchange rate risk that international investments are exposed to. Finally, we have discussed how movements in exchange rates affect international trade flows (imports and exports) and capital flows.

The following points, among others, are made in this chapter:

- Measured by average daily turnover, the foreign exchange market is by far the largest financial market in the world. It has important effects, either directly or indirectly, on the pricing and flows in all other financial markets.
- There is a wide diversity of global FX market participants that have a wide variety of motives for entering into foreign exchange transactions.
- Individual currencies are usually referred to by standardized three-character codes. These currency codes can also be used to define exchange rates (the price of one currency in terms of another). There are a variety of exchange rate quoting conventions commonly used.
- A direct currency quote takes the domestic currency as the price currency and the foreign currency as the base currency (i.e., $S_{d/f}$). An indirect quote uses the domestic currency as the base currency (i.e., $S_{f/d}$). To convert between direct and indirect quotes, the inverse (reciprocal) is used. Professional FX markets use standardized conventions for how the exchange rate for specific currency pairs will be quoted.
- Currencies trade in foreign exchange markets based on nominal exchange rates. An increase in the exchange rate, quoted in indirect terms, means that the domestic currency is appreciating versus the foreign currency, and a decrease in the exchange rate means the domestic currency is depreciating.
- The real exchange rate, defined as the nominal exchange rate multiplied by the ratio of price levels, measures the relative purchasing power of the currencies. An increase in the real exchange rate ($R_{d/f}$) implies a reduction in the relative purchasing power of the domestic currency.
- Given exchange rates for two currency pairs—A/B and A/C—we can compute the cross-rate (B/C) between currencies B and C. Depending on how the rates are quoted, this may require inversion of one of the quoted rates.
- Spot exchange rates are for immediate settlement (typically, T + 2), while forward exchange rates are for settlement at agreed-upon future dates. Forward rates can be used to manage foreign exchange risk exposures or can be combined with spot transactions to create FX swaps.
- The spot exchange rate, the forward exchange rate, and the domestic and foreign interest rates must jointly satisfy an arbitrage relationship that equates the investment return on two alternative but equivalent investments. Given the spot exchange rate and the foreign and domestic interest rates, the forward exchange rate must take the value that prevents riskless arbitrage.
- Forward rates are typically quoted in terms of forward (or swap) points. The swap points are added to the spot exchange rate in order to calculate the forward rate. Occasionally, forward rates are presented in terms of percentages relative to the spot rate.
- The base currency is said to be trading at a forward premium if the forward rate is above the spot rate (forward points are positive). Conversely, the base currency is said to be trading at a forward discount if the forward rate is below the spot rate (forward points are negative).
- The currency with the higher interest rate will trade at a forward discount, and the currency with the lower interest rate will trade at a forward premium.
- Swap points are proportional to the spot exchange rate and to the interest rate differential and approximately proportional to the term of the forward contract.
- Empirical studies suggest that forward exchange rates may be unbiased predictors of future spot rates, but the margin of error on such forecasts is too large for them to be used in practice as a guide to managing exchange rate exposures. FX markets are too complex and

too intertwined with other global financial markets to be adequately characterized by a single variable, such as the interest rate differential.

- Virtually every exchange rate is managed to some degree by central banks. The policy framework that each central bank adopts is called an exchange rate regime. These regimes range from using another country's currency (dollarization) to letting the market determine the exchange rate (independent float). In practice, most regimes fall in between these extremes. The type of exchange rate regime used varies widely among countries and over time.
- An ideal currency regime would have three properties: (1) the exchange rate between any two currencies would be credibly fixed; (2) all currencies would be fully convertible; and (3) each country would be able to undertake fully independent monetary policy in pursuit of domestic objectives, such as growth and inflation targets. However, these conditions are inconsistent. In particular, a fixed exchange rate and unfettered capital flows severely limit a country's ability to undertake independent monetary policy. Hence, there cannot be an ideal currency regime.
- The IMF identifies the following types of regimes: dollarization, monetary union, currency board, fixed parity, target zone, crawling peg, crawling band, managed float, and independent float. Most major currencies traded in FX markets are freely floating, albeit subject to occasional central bank intervention.
- A trade surplus must be matched by a corresponding deficit in the capital account, and a trade deficit must be matched by a capital account surplus. Any factor that affects the trade balance must have an equal and opposite impact on the capital account, and vice versa.
- A trade surplus reflects an excess of domestic saving (including the government fiscal balance) over investment spending. A trade deficit indicates that the country invests more than it saves and must finance the excess by borrowing from foreigners or selling assets to foreigners.
- The impact of the exchange rate on trade and capital flows can be analyzed from two perspectives. The elasticities approach focuses on the effect of changing the relative prices of domestic and foreign goods. This approach highlights changes in the composition of spending. The absorption approach focuses on the impact of exchange rates on aggregate expenditure/saving decisions.
- The elasticities approach leads to the Marshall–Lerner condition, which describes combinations of export and import demand elasticities such that depreciation of the domestic currency will move the trade balance toward surplus, and domestic currency appreciation will move the trade balance toward deficit.
- The idea underlying the Marshall–Lerner condition is that demand for imports and exports must be sufficiently price sensitive that an increase in the relative prices of imports increases the difference between export receipts and import expenditures.
- In order to move the trade balance toward surplus, a change in the exchange rate must decrease domestic expenditure (also called absorption) relative to income; to move the trade balance toward deficit, a change in the exchange rate must increase domestic expenditure. Equivalently, it must increase (or decrease) domestic saving relative to domestic investment.
- If there is excess capacity in the economy, then currency depreciation can increase output/income by switching demand toward domestically produced goods and services. Because some of the additional income will be saved, income rises relative to expenditure and the trade balance improves.
- If the economy is at full employment, then currency depreciation must reduce domestic expenditure in order to improve the trade balance. The main mechanism is a wealth effect: A weaker currency reduces the purchasing power of domestic-currency-denominated assets (including the present value of current and future earned income), and households respond by reducing expenditure and increasing saving.

PRACTICE PROBLEMS[22]

1. An exchange rate:
   A. is most commonly quoted in real terms.
   B. is the price of one currency in terms of another.
   C. between two currencies ensures that they are fully convertible.

2. A decrease in the real exchange rate (quoted in terms of domestic currency per unit of foreign currency) is *most likely* to be associated with an increase in which of the following?
   A. Foreign price level
   B. Domestic price level
   C. Nominal exchange rate

3. In order to minimize the foreign exchange exposure on a euro-denominated receivable due from a German company in 100 days, a British company would *most likely* initiate a:
   A. spot transaction.
   B. forward contract.
   C. real exchange rate contract.

4. Which of the following counterparties is *most likely* to be considered a sell-side foreign exchange market participant?
   A. A large corporation that borrows in foreign currencies
   B. A sovereign wealth fund that influences cross-border capital flows
   C. A multinational bank that trades foreign exchange with its diverse client base

5. What will be the effect on a direct exchange rate quote if the domestic currency appreciates?
   A. Increase
   B. Decrease
   C. No change

6. An executive from Switzerland checked into a hotel room in Spain and was told by the hotel manager that 1 EUR will buy 1.2983 CHF. From the executive's perspective, an indirect exchange rate quote would be:
   A. 0.7702 EUR per CHF.
   B. 0.7702 CHF per EUR.
   C. 1.2983 EUR per CHF.

7. Over the past month, the Swiss franc (CHF) has depreciated 12 percent against pound sterling (GBP). How much has the pound sterling appreciated against the Swiss franc?
   A. Exactly 12 percent
   B. Less than 12 percent
   C. More than 12 percent

[22]These practice problems were developed by Ryan Fuhrmann, CFA (Westfield, Indiana, USA).

8. An exchange rate between two currencies has increased to 1.4500. If the base currency has appreciated by 8 percent against the price currency, the initial exchange rate between the two currencies was *closest* to:
   A. 1.3340.
   B. 1.3426.
   C. 1.5660.

The following information relates to Questions 9 and 10.

A dealer provides the following quotes:

Ratio	Spot Rate
CNY/HKD	0.8422
CNY/ZAR	0.9149
CNY/SEK	1.0218

9. The spot ZAR/HKD cross-rate is *closest* to:
   A. 0.9205.
   B. 1.0864.
   C. 1.2978.

10. Another dealer is quoting the ZAR/SEK cross-rate at 1.1210. The arbitrage profit that can be earned is *closest* to:
    A. ZAR3,671 per million SEK traded.
    B. SEK4,200 per million ZAR traded.
    C. ZAR4,200 per million SEK traded.

11. A BRL/MXN spot rate is listed by a dealer at 0.1378. The six-month forward rate is 0.14193. The six-month forward points are *closest* to:
    A. −41.3.
    B. +41.3.
    C. +299.7.

12. A three-month forward exchange rate in CAD/USD is listed by a dealer at 1.0123. The dealer also quotes three-month forward points as a percentage at 6.8 percent. The CAD/USD spot rate is *closest* to:
    A. 0.9478.
    B. 1.0550.
    C. 1.0862.

13. If the base currency in a forward exchange rate quote is trading at a forward discount, which of the following statements is *most* accurate?
    A. The forward points will be positive.
    B. The forward percentage will be negative.
    C. The base currency is expected to appreciate versus the price currency.

14. A forward premium indicates:
    A. an expected increase in demand for the base currency.
    B. that the interest rate is higher in the base currency than in the price currency.
    C. that the interest rate is higher in the price currency than in the base currency.

15. The JPY/AUD spot exchange rate is 82.42, the JPY interest rate is 0.15 percent, and the AUD interest rate is 4.95 percent. If the interest rates are quoted on the basis of a 360-day year, the 90-day forward points in JPY/AUD would be *closest* to:
    A. −377.0.
    B. −97.7.
    C. 98.9.

16. Which of the following is *not* a condition of an ideal currency regime?
    A. Fully convertible currencies
    B. Fully independent monetary policy
    C. Independently floating exchange rates

17. In practice, both a fixed parity regime and a target zone regime allow the exchange rate to float within a band around the parity level. The *most likely* rationale for the band is that the band allows the monetary authority to:
    A. be less active in the currency market.
    B. earn a spread on its currency transactions.
    C. exercise more discretion in monetary policy.

18. A fixed exchange rate regime in which the monetary authority is legally required to hold foreign exchange reserves backing 100 percent of its domestic currency issuance is best described as:
    A. dollarization.
    B. a currency board.
    C. a monetary union.

19. A country with a trade deficit will *most likely*:
    A. have an offsetting capital account surplus.
    B. save enough to fund its investment spending.
    C. buy assets from foreigners to fund the imbalance.

20. A large industrialized country has recently devalued its currency in an attempt to correct a persistent trade deficit. Which of the following domestic industries is *most likely* to benefit from the devaluation?
    A. Luxury cars
    B. Branded prescription drugs
    C. Restaurants and live entertainment venues

21. A country with a persistent trade surplus is being pressured to let its currency appreciate. Which of the following *best* describes the adjustment that must occur if currency appreciation is to be effective in reducing the trade surplus?
    A. Domestic investment must decline relative to saving.
    B. Foreigners must increase investment relative to saving.
    C. Global capital flows must shift toward the domestic market.

CHAPTER 10

CURRENCY EXCHANGE RATES: DETERMINATION AND FORECASTING

Michael R. Rosenberg

William A. Barker, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Calculate and interpret the bid/ask spread on a spot or forward foreign currency quotation and describe the factors that affect the bid/offer spread.
- Identify a triangular arbitrage opportunity and calculate its profit, given the bid/offer quotations for three currencies.
- Distinguish between spot and forward rates and calculate the forward premium or discount for a given currency.
- Calculate the mark-to-market value of a forward contract.
- Explain international parity relations—covered and uncovered interest rate parity, purchasing power parity, and the international Fisher effect.
- Describe relationships among the international parity conditions.
- Evaluate the use of the current spot rate, the forward rate, purchasing power parity, and uncovered interest parity to forecast future spot exchange rates.
- Explain approaches to assessing the long-run fair value of an exchange rate.
- Describe the carry trade and its relationship to uncovered interest rate parity, and calculate the profit from such a strategy.
- Explain how flows in the balance of payment accounts affect currency exchange rates.
- Describe the Mundell–Fleming model, the monetary approach, and the asset market (portfolio balance) approach to exchange rate determination.
- Forecast the direction of the expected change in an exchange rate based on balance of payment, Mundell–Fleming, monetary, and asset market approaches to exchange rate determination.

- Explain the potential impacts of monetary and fiscal policies on exchange rates.
- Describe the objectives and effectiveness of central bank intervention and capital controls.
- Describe warning signs of a currency crisis.
- Describe the use of technical analysis in forecasting exchange rates.

1. INTRODUCTION

Niels Bohr, the famous Danish physicist, once joked that "prediction is very difficult, especially about the future." No words could better express the difficulties associated with exchange rate forecasting. As anyone involved in the business of currency forecasting can attest, it can be a humbling experience. Alan Greenspan, former U.S. Federal Reserve chairman, famously noted that "having endeavored to forecast exchange rates for more than half a century, I have understandably developed significant humility about my ability in this area." Bill Gross, co-chief investment officer at PIMCO, once commented that "if you think writing about the fortunes of the stock market is tricky, try getting your arms around currencies."

The purpose of this chapter is to provide readers with tools that will better enable them to get their "arms around currencies." Economists have developed a wide range of theories to explain how exchange rates are determined. This chapter discusses the main theories in detail—starting with the basic international parity building blocks, moving on to long-run equilibrium models, and then digging deeper into the important medium-term cyclical drivers such as monetary policy, fiscal policy, current account balances, and capital flow trends.

In addition, the chapter reviews the empirical evidence to assess how well our theoretical models stack up in practice. In short, the empirical evidence shows that real-world exchange rates have large and persistent deviations from their theoretical long-run equilibrium values. Indeed, most studies conclude that for short- and medium-term horizons, up to perhaps a few years, a random walk characterizes exchange rate movements better than most fundamentals-based exchange rate models. Most studies find that models that work well in one period fail in others. Most studies also find that models that work for one set of exchange rates fail to work for others.

One of the key reasons why fundamentals-based models perform so poorly in the short run is that changes in fundamental economic variables such as relative money supplies, interest rates, inflation rates, economic growth rates, and current account balances simply do not exhibit anywhere near the variability that exchange rates exhibit on a monthly or quarterly basis. As a result, researchers are often unable to unearth any significant contemporaneous relationship between changes in macroeconomic variables and changes in exchange rates over short- and medium-term horizons. Exchange rates may move in the direction suggested by economic fundamentals in the long run, but the often chaotic behavior of exchange rates over short- and medium-term periods is capable of generating so much noise that it tends to obscure any discernible relationship between macroeconomic time series and short- and medium-term exchange rate movements.

Given the shortcomings of most fundamentals-based models, currency strategists and market participants often have felt compelled to turn to non-fundamentals-based forecasting tools to get a better handle on shorter-run exchange rate trends. Such forecasting tools include technically based trend-following trading rules and order flow, sentiment, and positioning indicators. Unfortunately, recent studies show that the predictive value of these forecasting tools is either mixed or nonexistent.

On a more positive note, there is evidence that certain systematic foreign exchange investment strategies have rewarded currency managers with relatively high excess returns over fairly long periods of time (that is, in excess of the risk-free rate of return). One such strategy that has attracted a lot of interest among international investors is the so-called foreign exchange (FX) carry trade. FX carry trades entail going long a basket of high-yielding currencies and simultaneously going short a basket of low-yielding currencies. Although the empirical evidence suggests that the excess returns on this strategy have been fairly attractive, investors need to be mindful that carry trades are prone to crash when market conditions become volatile. Hence, investors need to overlay simple carry trade strategies with well-thought-out risk management systems to help protect against downside risks.

This chapter discusses in detail various approaches that economists and market strategists have devised on how to best position oneself in the currency markets. The reader should develop a general understanding of the fundamental and technical forces that affect exchange rates over short-, medium- and long-run horizons. At the same time, the reader should develop an appreciation of the issues that one is likely to face in devising a successful and profitable exchange rate forecasting and/or trading strategy.

The chapter proceeds as follows:

- Section 2 reviews the basic concepts of the foreign exchange market covered in preceding chapters and expands this previous coverage to incorporate more material on bid/offer spreads.
- In Section 3, we begin to examine determinants of exchange rates, starting with longer-term interrelationships among exchange rates, interest rates, and inflation rates embodied in the international parity conditions. Despite their empirical shortcomings, these parity conditions form the key building blocks for many long-run exchange rate models. We then examine alternative approaches for determining long-term "fair value" for a currency, and we use this expanded approach to derive an explanatory framework that shows how medium-term factors can cause observed exchange rates to fluctuate around a path to long-term equilibrium.
- Section 4 examines the FX carry trade, a profitable trading strategy that exploits exchange rate deviations from uncovered interest rate parity (one of the international parity conditions).
- Section 5 examines the relationship between a country's exchange rate and its balance of payments using the analytical framework developed earlier in Section 3.
- In Section 6, we examine how monetary and fiscal policies can *indirectly* influence exchange rates by influencing the various factors described in our exchange rate model from Section 3.
- Section 7 examines *direct* public-sector actions in foreign exchange markets, either through capital controls or by foreign exchange market intervention (buying and selling currencies for policy purposes).
- Section 8 examines historical episodes of currency crises and some leading indicators that may signal increased likelihood of a crisis.
- Having examined the longer- and medium-term influences on exchange rates, in Section 9 we examine some of the tools for predicting exchange rate movements over shorter time horizons: technical analysis; order flow, sentiment, and positioning measures; and indicators derived from options and futures markets.

A final section summarizes the key points of the chapter, followed by an appendix of currency codes and a set of practice problems.

2. FOREIGN EXCHANGE MARKET CONCEPTS

We begin with a brief review of some of the basic conventions of the FX market that were covered in the preceding chapters.

An exchange rate is the price of the *base* currency expressed in terms of the *price* currency. For example, a USD/EUR rate of 1.3650 means the euro, the base currency, costs 1.3650 U.S. dollars (an appendix defines the three-letter currency codes used in this chapter). The exact notation used to represent exchange rates can vary widely between sources, and occasionally the same exchange rate notation will be used by different sources to mean completely different things. *The reader should be aware that the notation used here may not be the same as that encountered elsewhere*. To avoid confusion, this chapter will identify exchange rates using the convention of "P/B" referring to the price of the base currency "B" expressed in terms of the price currency "P."[1]

The spot exchange rate is usually for settlement on the second business day after the trade date, referred to as T + 2 settlement.[2] In foreign exchange markets—as in other financial markets—market participants confront a two-sided price in the form of a bid price and an offer price (also called an ask price) quoted by potential counterparties. The bid price is the price, defined in terms of the price currency, at which the counterparty providing a two-sided price quote is willing to buy one unit of the base currency. Similarly, the offer price is the price, in terms of the price currency, at which that counterparty is willing to sell one unit of the base currency. For example, given a price request from a client, a dealer might quote a two-sided price on the spot USD/EUR exchange rate of 1.3648/1.3652. This means that the dealer is willing to pay USD1.3648 to buy one euro and that the dealer will sell one euro for USD1.3652.

There are two points to bear in mind about bid/offer quotes:

1. *The offer price is always higher than the bid price*. The bid/offer spread—the difference between the offer price and the bid price—is the compensation that counterparties seek for providing foreign exchange to other market participants.
2. The counterparty in the transaction who inquires for a two-sided price quote will have the option (but not the obligation) to deal at either the bid (to sell the base currency) or offer (to buy the base currency) price quoted to the inquirer by the dealer. The inquirer can pass on the price quote, but if the inquirer deals, the jargon in the market is that this counterparty has either "hit the bid" or "paid the offer." To determine whether the bid or offer side of the market should be used in terms of describing a foreign exchange

[1]Notation is generally not standardized in global foreign exchange markets, and there are several common ways of expressing the same currency pair (e.g., JPY/USD, USD:JPY, $/¥). What is common in FX markets, however, is the concept of a base and price currency when setting exchange rate prices. Later in the chapter, we will sometimes switch to discussing a "domestic" currency and a "foreign" currency quoted as foreign/domestic (*f/d*). This will be only an illustrative device for more easily explaining various theoretical concepts. The reader should be aware that describing currency pairs in terms of "foreign" and "domestic" currencies is not done in professional FX markets. This is because what is the "foreign" and what is the "domestic" currency depend on where one is located, which can lead to confusion. For instance, what is "foreign" and what is "domestic" for a Middle Eastern investor trading CHF against GBP with the New York branch of a European bank, with the trade ultimately booked at the bank's headquarters in Paris?

[2]The exception among the major currencies is CAD/USD, for which standard spot settlement is T + 1.

transaction, one should first determine which currency is the base currency in the currency quote and then determine whether the base currency is being sold (hit the bid) or bought (pay the offer).

We will distinguish here between the bid/offer pricing a *client receives from a dealer* and the pricing a *dealer receives from the interbank market*. Dealers buy and sell foreign exchange among themselves in what is called the interbank market.[3] This global network for exchanging currencies among professional market participants allows dealers to adjust their inventories and risk positions, distribute foreign currencies to end users who need them, and transfer foreign exchange rate risk to market participants who are willing to bear it. The interbank market is typically for dealing sizes of at least one million units of the base currency. Of course, the dealing amount can be larger than a million units; indeed, interbank market trades generally are measured in terms of multiples of a million units of the base currency.

The bid/offer spread a dealer provides to most clients typically is slightly wider than the bid/offer spread observed in the interbank market. For example, if the quote in the interbank USD/EUR spot market is 1.3649/1.3651 (two pips wide), the dealer might quote a client a bid/offer of 1.3648/1.3652 (four pips wide) for a spot USD/EUR transaction. When the dealer buys the base currency from a client, the dealer typically wants to turn around and sell the base currency in the interbank market, and when the dealer sells the base currency to a client, the dealer typically wants to turn around and buy the base currency in the interbank market. This offsetting transaction allows the dealer to both get out of the risk exposure assumed by providing a two-sided price to the client and also make a profit. Continuing our example, suppose the dealer's client hits the dealer's bid and sells EUR to the dealer for USD1.3648. The dealer is now long EUR (and short USD) and wants to cover this position in the interbank market. To do this, the dealer sells the EUR in the interbank market by hitting the interbank bid. As a result, the dealer *bought* EUR from the client at USD1.3648 and then *sold* the EUR in the interbank for USD1.3649. This gives the dealer a profit of USD0.0001 (one pip) for every EUR transacted. This one pip translates into a profit of USD100 per EUR1 million bought from the client. If, instead of hitting his bid, the client paid the offer (1.3652), then the dealer could pay the offer in the interbank market (1.3651), earning a profit of one pip.

The size of the bid/offer spread, in pips, quoted to dealers' clients in the FX market can vary widely across exchange rates and is not constant over time, even for a single exchange rate. The size of this bid/offer spread depends primarily on three factors: the bid/offer spread in the interbank foreign exchange market for the two currencies involved, the size of the transaction, and the relationship between the dealer and the client. We examine each factor in turn.

The size of the bid/offer spread quoted in the interbank market depends on the liquidity in this market, which in turn depends on several factors, including the following:

- *The currency pair involved.* Market participation is greater for some currency pairs than others. Liquidity in the major currency pairs—for example, USD/EUR, JPY/USD, or USD/GBP—can be considerable. These markets are almost always deep with bids and offers from market participants around the world. In other currency pairs, particularly some

[3]Although we refer to this as the interbank market, many nonbank entities can now access this market. These nonbank entities include institutional asset managers, hedge funds, and other large, sophisticated market participants. Detailed discussion of this topic is beyond the scope of this chapter.

of the more obscure currency cross-rates (for example, MXN/CHF), market participation is much thinner and consequently the bid/offer spread in the interbank market will be wider.

- *The time of day.* The interbank FX markets are most liquid when the major FX trading centers are open. Business hours in London and New York—the two largest FX trading centers—overlap from approximately 8:00 a.m. to 11:00 a.m. New York time. The interbank FX market for most currency pairs is typically most liquid during these hours. After London closes, liquidity is thinner through the New York afternoon. The Asian session starts when dealers in Tokyo, Singapore, and Hong Kong open for business, typically by 7:00 p.m. New York time. For most currency pairs, however, the Asian session is not as liquid as the London and New York sessions. Although FX markets are open 24 hours a day on business days, liquidity in interbank markets can be very thin between the time New York closes and the time Asia opens, because Sydney, Australia, tends to be the only active trading center during these hours. For reference, the following chart shows a 24-hour period from midnight (00:00) to midnight (24:00) London time, corresponding standard times in Tokyo and New York, and the *approximate* hours of the most liquid trading periods in each market.

Standard Time and Approximate FX Trading Hours in Major Markets: Midnight to Midnight London Time

Tokyo	09:00	13:00	17:00	21:00	01:00 Day + 1	05:00 Day + 1	09:00 Day + 1
London	00:00	04:00	08:00	12:00	16:00	20:00	24:00
New York	19:00 Day − 1	23:00 Day − 1	03:00	07:00	11:00	15:00	19:00

- *Market volatility.* As in any financial market, when major market participants have greater uncertainty about the factors influencing market pricing, they will attempt to reduce their risk exposures and/or charge a higher price for taking on risk. In the FX market, this response implies wider bid/offer spreads in both the interbank and broader markets. Geopolitical events (e.g., war, civil strife); market crashes; and major data releases (e.g., U.S. nonfarm payrolls) are among the factors that influence spreads and liquidity.

The size of the transaction can also affect the bid/offer spread shown by a dealer to clients. Typically, the larger the transaction, the further away from the current spot exchange rate the dealing price will be. Hence, a client who asks a dealer for a two-sided spot CAD/USD price on, for example, USD50 million will be shown a wider bid/offer spread than a client who asks for a price on USD1 million. The wider spread reflects the greater difficulty the dealer faces in laying off the foreign exchange risk of the position in the interbank FX market.[4]

[4]Smaller dealing sizes can also affect the bid/offer quote shown to clients. Retail quotes are typically for dealing sizes less than one million units of the base currency and can range all the way down to foreign exchange transactions conducted by individuals (for example, going to a local bank in order to purchase foreign currency for an overseas holiday). Depending on the dealing venue, the bid/offer spreads for these retail transactions can be very large compared with the interbank market. Whereas the bid/offer spread in the interbank market for most liquid currency pairs can be a pip or two, going to the teller window at a local bank branch or using a credit card to conduct FX transactions can result in a retail bid/offer spread of several hundreds of pips.

The relationship between the dealer and the client can also affect the size of the bid/offer spread shown to clients. For many clients, the spot foreign exchange business is only one business service among many that a dealer provides to the client. For example, the dealer firm might also transact in bond and/or equity securities with the same client. In a competitive business environment, in order to win the client's business for these other services, the dealer might provide a tighter (i.e., smaller) bid/offer spot exchange rate quote. The dealer might also give tighter bid/offer quotes in order to win repeat FX business. A client's credit risk can also be a factor. A client with a poor credit profile may be quoted a wider bid/offer spread than one with good credit. Given the short settlement cycle for spot FX transactions (typically two business days), however, credit risk is not the most important factor in determining the client's bid/offer spread on spot exchange rates.

2.1. Arbitrage Constraints on Spot Exchange Rate Quotes

The bid/offer quotes a dealer shows in the interbank FX market must respect two arbitrage constraints; otherwise the dealer creates riskless arbitrage opportunities for other interbank market participants.[5]

First, the bid shown by a dealer in the interbank market cannot be higher than the current interbank offer, and the offer shown by a dealer cannot be lower than the current interbank bid. If the bid/offer quotes shown by a dealer are inconsistent with the interbank market quotes, other market participants will buy from the cheaper source and sell to the more expensive source. This arbitrage will eventually bring the two prices back into line. For example, suppose that the current spot USD/EUR price in the interbank market is 1.3649/1.3651. If a dealer showed a misaligned price quote of 1.3652/1.3654, then other market participants would pay the offer in the interbank market, *buying* EUR at a price of USD1.3651, and then *sell* the EUR to the dealer by hitting the dealer's bid at USD1.3652—thereby making a riskless profit of one pip on the trade. This arbitrage would continue as long as the dealer's bid/offer quote violated the arbitrage constraint.

Second, the cross-rate bids posted by a dealer must be lower than the implied cross-rate offers available in the interbank market, and the dealer's offers must be higher than the available bids. Recall that given exchange rate quotes for the currency pairs A/B and C/B, we can back out the implied cross-rate of A/C, and that this implied cross-rate A/C must be consistent with the A/B and C/B rates. This again reflects the basic principle of arbitrage: If identical financial products are priced differently, then market participants will buy the cheaper one and sell the more expensive one until the price difference is eliminated. In the context of FX cross-rates, there are two ways to trade currency A against currency C: (1) using the cross-rate A/C or (2) using the A/B and C/B rates. Because, in the end, both methods involve selling (or buying) currency C in order to buy (or sell) currency A, the exchange rate using these two approaches must be consistent.

To illustrate this **triangular arbitrage** among three currencies, suppose that the interbank market bid/offer in USD/EUR is 1.3649/1.3651 and that the bid/offer in JPY/USD is 76.64/76.66. We need to use these two interbank bid/offer quotes to calculate the market-implied bid/offer quote on the JPY/EUR cross-rate.

[5]We will confine our attention to the interbank FX market because arbitrage presumes the ability to deal simultaneously with different market participants and in different markets, to be able to access wholesale bid/offer quotes, and to have the market sophistication to spot arbitrage opportunities. These criteria are typically limited to interbank market participants—that is, the professional FX market.

Let us begin by considering the transactions required to *sell* JPY and *buy* EUR, going through the JPY/USD and USD/EUR currency pairs. We can view this intuitively as follows:

Sell JPY Buy EUR	=	Sell JPY Buy USD	then	Sell USD Buy EUR

Note that the "Buy USD" and "Sell USD" in the expressions on the right-hand side of the equal sign will cancel out to give the JPY/EUR cross-rate. In equation form, we can represent this relationship as follows:

$$\left(\frac{\text{JPY}}{\text{EUR}}\right) = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)$$

Now, let's take account of the bid/offer rates in order to do the JPY/EUR calculation.

1. The left-hand side of the equal sign is "Sell JPY, Buy EUR." In the JPY/EUR price quote, the EUR is the base currency, and buying it means paying the offer; that is, we will be calculating the *offer* rate in JPY/EUR.
2. The first term on the right-hand side of the equal sign is "Sell JPY, Buy USD." Because the USD is the base currency in the JPY/USD quote, buying it means paying the *offer*.
3. The second term on the right-hand side of the equal sign is "Sell USD, Buy EUR." Because the EUR is the base currency in the USD/EUR price quote, buying it means paying the *offer*.

Combining all of this together conceptually and putting in the relevant offer rates leads to:

$$\left(\frac{\text{JPY}}{\text{EUR}}\right)_{\text{offer}} = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)_{\text{offer}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{offer}} = 76.66 \times 1.3651 = 104.65$$

Perhaps not surprisingly, calculating the implied JPY/EUR *bid* rate uses the same process as before (now using "Buy JPY, Sell EUR" for the left-hand side of the equation). This leads to:

$$\left(\frac{\text{JPY}}{\text{EUR}}\right)_{\text{bid}} = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)_{\text{bid}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{bid}} = 76.64 \times 1.3649 = 104.61$$

As one would expect, the implied cross-rate bid is less than the offer: 104.61/104.65.

This simple formula seems relatively straightforward: To get the implied *bid* cross-rate, simply multiply the *bid* rates for the other two currencies. One must be cautious about the quoting conventions used for the currencies, however, because this simple formula is *not always the case*. Depending on quoting conventions, it may be necessary to *invert* one of the quotes in order to complete the calculation.

This is best illustrated with an example. Consider the case of calculating the implied GBP/EUR cross-rate given USD/GBP and USD/EUR quotes. In this case, simple multiplication will not work, and we have to invert the USD/GBP quote:

$$\frac{\text{GBP}}{\text{EUR}} \neq \left(\frac{\text{USD}}{\text{GBP}}\right)\left(\frac{\text{USD}}{\text{EUR}}\right)$$

Instead, we have:

$$\frac{\text{GBP}}{\text{EUR}} = \left(\frac{\text{GBP}}{\cancel{\text{USD}}}\right)\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right) = \left(\frac{\text{USD}}{\text{GBP}}\right)^{-1}\left(\frac{\text{USD}}{\text{EUR}}\right)$$

The implied *bid* rate from this expression is the rate at which the customer can "buy the GBP, sell the EUR," because the EUR is the base currency in the GBP/EUR quote. Conceptually, calculating the implied *bid* rate for GBP/EUR proceeds by:

Buy GBP Sell EUR	=	Buy GBP Sell USD	then	Buy USD Sell EUR

As we saw above, however, we need to *invert* USD/GBP in order to get GBP/USD, the first expression to the right of the equal sign. Let's assume that the bid/offer in USD/GBP is 1.5644/1.5646. Inverting this expression gives the bid/offer in GBP/USD as equal to 0.63914/0.63922. Note that the bid must always be smaller than the offer. Hence, to get the GBP/USD *bid*, we are using the inverse of the USD/GBP *offer*. Similarly, the GBP/USD *offer* is calculated by inverting the USD/GBP *bid*. (Note that we extended the calculated GBP/USD quotes to five decimal places to avoid truncation errors in subsequent calculations.)

As before, let's use 1.3649/1.3651 as the bid/offer in USD/EUR. We then combine all this together conceptually and mathematically to calculate the GBP/EUR *bid* rate:

$$\left(\frac{\text{GBP}}{\text{EUR}}\right)_{\text{bid}} = \left(\frac{\text{GBP}}{\cancel{\text{USD}}}\right)_{\text{bid}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{bid}} = 0.63914 \times 1.3649 = 0.8724$$

Similarly, for the implied GBP/EUR *offer* rate:

$$\left(\frac{\text{GBP}}{\text{EUR}}\right)_{\text{offer}} = \left(\frac{\text{GBP}}{\cancel{\text{USD}}}\right)_{\text{offer}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{offer}} = 0.63922 \times 1.3651 = 0.8726$$

Note that the implied *bid* rate is less than the implied *offer* rate, as it must be to prevent arbitrage.

We conclude this section on arbitrage constraints with some simple observations:

- The arbitrage constraint on implied cross-rates is similar to that for spot rates (posted bid rates cannot be higher than the market's offer; posted offer rates cannot be lower than the market's bid). The only difference is that this second arbitrage constraint is applied *across* currency pairs instead of involving a *single* currency pair.
- In reality, violations of these arbitrage constraints almost never occur. Both human traders and automatic trading algorithms are constantly on alert for any pricing inefficiencies and will arbitrage them away almost instantly.
- Market participants never have to calculate cross-rates *manually* because electronic dealing machines (which are essentially just specialized computers) will automatically calculate cross bid/offer rates given any two underlying bid/offer rates.

EXAMPLE 10-1 Bid/Offer Rates

The following are spot rate quotes in the interbank market:

USD/EUR	1.4559/1.4561
JPY/USD	81.87/81.89
CAD/USD	0.9544/0.9546
SEK/USD	6.8739/6.8741

1. What is the bid/offer on the SEK/EUR cross-rate implied by the interbank market?
   A. 0.2118/0.2119
   B. 4.7209/4.7214
   C. 10.0077/10.0094
2. What is the bid/offer on the JPY/CAD cross-rate implied by the interbank market?
   A. 78.13/78.17
   B. 85.781/85.785
   C. 85.76/85.80
3. If a dealer quoted a bid/offer rate of 85.73/85.75 in JPY/CAD, then a triangular arbitrage would involve buying:
   A. CAD in the interbank market and selling it to the dealer, for a profit of JPY 0.01 per CAD.
   B. JPY from the dealer and selling it in the interbank market, for a profit of CAD 0.01 per JPY.
   C. CAD from the dealer and selling it in the interbank market, for a profit of JPY 0.01 per CAD.
4. If a dealer quoted a bid/offer of 85.74/85.81 in JPY/CAD, then you could:
   A. not make any arbitrage profits.
   B. make arbitrage profits by buying JPY from the dealer and selling it in the interbank market.
   C. make arbitrage profits by buying CAD from the dealer and selling it in the interbank market.
5. A market participant is considering the following transactions:

Transaction 1: Buy CAD 100 million against the USD at 15:30 London time.
Transaction 2: Sell CAD 100 million against the KRW at 21:30 London time.
Transaction 3: Sell CAD 10 million against the USD at 15:30 London time.

Given the proposed transactions, what is the *most likely* ranking of the bid/ask spreads, from tightest to widest, under normal market conditions?

A. Transaction 1, 2, 3
B. Transaction 2, 1, 3
C. Transaction 3, 1, 2

Solution to 1: C is correct. Using the intuitive equation-based approach,

$$\frac{\text{SEK}}{\text{EUR}} = \frac{\text{SEK}}{\text{USD}} \times \frac{\text{USD}}{\text{EUR}}$$

Hence, to calculate the SEK/EUR bid/offer rate, we multiply the SEK/USD and USD/EUR bid and offer rates to get:

Bid: 10.0077 (= 6.8739 × 1.4559)

Offer: 10.0094 (= 6.8741 × 1.4561)

Solution to 2: C is correct. Using the intuitive equation-based approach,

$$\frac{\text{JPY}}{\text{CAD}} = \frac{\text{JPY}}{\text{USD}} \times \left(\frac{\text{CAD}}{\text{USD}}\right)^{-1} = \frac{\text{JPY}}{\text{~~USD~~}} \times \frac{\text{~~USD~~}}{\text{CAD}}$$

This equation shows that we have to invert the CAD/USD quotes to get the USD/CAD bid/offer rates of 1.04756/1.04778. That is, given the CAD/USD quotes of 0.9544/0.9546, take the inverse of each and interchange bid and offer, such that the USD/CAD quotes are (1/0.9546)/(1/0.9544) or 1.04756/1.04778. Multiplying the JPY/USD and USD/CAD bid/offer rates then leads to:

Bid: 85.76 = 81.87 × 1.04756

Offer: 85.80 = 81.89 × 1.04778

Solution to 3: C is correct. The interbank-implied cross-rate for JPY/CAD is 85.76/85.80 (the answer to Question 2). Hence, the dealer is posting an offer rate to sell the CAD (the base currency in the quote) too cheaply, at a rate below the interbank bid rate (85.75 versus 85.76, respectively). Hence triangular arbitrage would involve buying CAD from the dealer (paying the dealer's offer) and selling CAD in the interbank market (hitting the interbank bid) for a profit of JPY 0.01 (85.76 – 85.75) per CAD transacted.

Solution to 4: A is correct. The arbitrage relationship is not violated: The dealer's bid is not above the interbank market's offer, and the offer is not below the bid. The implied interbank cross-rate for JPY/CAD is 85.76/85.80 (the solution to Question 2).

Solution to 5: C is correct. The CAD/USD currency pair is most liquid when New York and London are both in their most liquid trading periods at the same time (approximately 8:00 a.m. to 11:00 a.m. New York time, or about 13:00 to 16:00 London time). Transaction 3 is for a smaller amount than transaction 1. Transaction 2 is for a less liquid currency pair (KRW/CAD is traded less than CAD/USD) and occurs outside of normal dealing hours in all three major centers (London, North America, and Asia); the transaction is also for a large amount.

2.2. Forward Markets

Outright forward contracts (often referred to simply as forwards) are agreements to exchange one currency for another on a future date at an exchange rate agreed on today. In contrast to spot rates, forward contracts are any exchange rate transactions that occur with settlement longer than the usual T + 2 settlement for spot delivery.

Forward exchange rates must satisfy an arbitrage relationship that equates the investment return on two alternative but equivalent investments. To simplify the explanation of this arbitrage relationship, and to focus on the intuition behind forward rate calculations, we will ignore the bid/offer spread on exchange rates and money market instruments. In addition, we will alter our exchange rate notation somewhat from price/base currency (P/B) to foreign/domestic currency (*f*/*d*). We will also assume that the domestic currency for an investor is the base currency in the standard exchange rate quotation. Using this *f*/*d* notation will make it easier to illustrate the choice an investor faces between domestic and foreign investments, as well as the arbitrage relationships that equate the returns on these investments when their risk characteristics are equal.

Consider an investor with one unit of domestic currency to invest for one year. The investor faces two alternatives:

1. One alternative is to invest cash for one year at the domestic risk-free rate (i_d). At the end of the year, the investment would be worth $(1 + i_d)$.
2. The other alternative is to convert the domestic currency to foreign currency at the spot rate of $S_{f/d}$ and invest for one year at the foreign risk-free rate (i_f). At the end of the period, the investor would have $S_{f/d}(1 + i_f)$ units of foreign currency. These funds then must be converted back to the investor's domestic currency. If the exchange rate to be used for this end-of-year conversion is set at the start of the period using a one-year forward contract, then the investor will have eliminated the foreign exchange risk associated with converting at an unknown future spot rate. Letting $F_{f/d}$ denote the forward rate, the investor would obtain $(1/F_{f/d})$ units of the domestic currency for each unit of foreign currency sold forward. Hence, in domestic currency, at the end of the year the investment would be worth $S_{f/d}(1 + i_f)\ (1/F_{f/d})$.

Because each of the two investments (A and B) is risk free, they must have the same return. Otherwise investors could earn a riskless arbitrage profit by selling (going short) one investment and investing in the other. In particular, investors could borrow in one currency, lend in the other, and use the spot and forward exchange markets to eliminate currency risk. Equating the returns on these two alternative investments—that is, putting investments A and B on opposite sides of the equal sign—leads to the following relationship:

$$(1 + i_d) = S_{f/d}(1 + i_f)\left(\frac{1}{F_{f/d}}\right)$$

To help see the intuition behind forward rate calculations, note that the right-hand side of the expression (for investment B) also shows the chronological order of this investment: Convert from domestic to foreign currency at the spot rate ($S_{f/d}$); invest this foreign currency amount at the foreign risk-free interest rate $(1 + i_f)$; and then, at maturity, convert the foreign currency investment proceeds back into the domestic currency using the forward rate $(1/F_{f/d})$.

For simplicity, we assumed a one-year horizon in the preceding example. However, the argument holds for any investment horizon. The risk-free assets used in this arbitrage

relationship are typically bank deposits quoted using the London Interbank Offered Rate (LIBOR) for the currencies involved. The day count convention for almost all LIBOR deposits is actual/360.[6] Incorporating this day count convention into our arbitrage formula leads to:

$$\left[1 + i_d\left(\frac{Actual}{360}\right)\right] = S_{f/d}\left[1 + i_f\left(\frac{Actual}{360}\right)\right]\left(\frac{1}{F_{f/d}}\right)$$

This equation can be rearranged to isolate the forward rate:

$$F_{f/d} = S_{f/d}\left[\frac{1 + i_f\left(\frac{Actual}{360}\right)}{1 + i_d\left(\frac{Actual}{360}\right)}\right] \tag{10-1}$$

This equation is known as **covered interest rate parity**. Our previous work shows that covered interest rate parity is based on an arbitrage relationship among risk-free interest rates and spot and forward exchange rates. Because of this arbitrage relationship between alternative investments, Equation 10-1 can also be described as saying that the covered (i.e., currency-hedged) interest rate differential between the two markets is zero.

The covered interest rate parity equation can also be rearranged to give an expression for the forward premium or discount:

$$F_{f/d} - S_{f/d} = S_{f/d}\left[\frac{\left(\frac{Actual}{360}\right)}{1 + i_d\left(\frac{Actual}{360}\right)}\right](i_f - i_d)$$

The domestic currency will trade at a forward premium ($F_{f/d} > S_{f/d}$) if, and only if, the foreign risk-free interest rate exceeds the domestic risk-free interest rate ($i_f > i_d$). The premium or discount is proportional to the spot exchange rate ($S_{f/d}$), proportional to the interest rate differential $i_f - i_d (i_f - i_d)$ between the markets, and approximately proportional to the time to maturity (actual/360).

Finally, although for simplicity's sake we have developed the **covered interest rate parity** equation (Equation 10-1) in terms of foreign and domestic currencies (using the notation *f/d*), this equation can equivalently be expressed in our more standard exchange rate quoting convention of price and base currencies (P/B):

$$F_{P/B} = S_{P/B}\left[\frac{1 + i_p\left(\frac{Actual}{360}\right)}{1 + i_B\left(\frac{Actual}{360}\right)}\right]$$

When dealing in professional FX markets, it is perhaps more useful to think of the covered interest rate parity equation and the calculation of forward rates in this P/B notation rather than foreign/domestic (*f/d*) notation. This is because domestic and foreign are relative

[6]This means that interest is calculated as if there are 360 days in a year. However, the actual number of days the funds are on deposit is used to calculate the interest payable. The main exception to the actual/360 day count convention is the GBP, for which the convention is actual/365. For the purposes of this chapter, we will use actual/360 consistently in order to avoid complication. In practice, however, one should confirm and apply the correct day count convention for each rate. Applying incorrect day counts could give the illusion of an arbitrage opportunity where none actually exists.

concepts that depend on where one is located, and because of the potential for confusion, these terms are not used for currency quotes in professional FX markets.

EXAMPLE 10-2 Calculating the Forward Premium and Discount

The following table shows the midmarket (i.e., average of the bid and offer) for the current CAD/AUD spot exchange rate as well as for AUD and CAD 270-day LIBOR (annualized):

Spot (CAD/AUD)	1.0145
270-day LIBOR (AUD)	4.87%
270-day LIBOR (CAD)	1.41%

The forward premium or discount for a 270-day forward contract for CAD/AUD would be *closest* to:

A. −0.0346.
B. −0.0254.
C. +0.0261.

Solution: B is correct. The equation to calculate the forward premium or discount is:

$$F_{P/B} - S_{P/B} = S_{P/B}\left[\frac{\left(\frac{Actual}{360}\right)}{1 + i_B\left(\frac{Actual}{360}\right)}\right](i_P - i_B)$$

Because AUD is the base currency in the CAD/AUD quote, putting in the information from the table leads to:

$$F_{P/B} - S_{P/B} = 1.0145\left[\frac{\left(\frac{270}{360}\right)}{1 + 0.0487\left(\frac{270}{360}\right)}\right](0.0141 - 0.0487) = -0.0254$$

In professional FX markets, forward exchange rates are typically quoted in terms of points. The points on a forward rate quote are simply the difference between the forward exchange rate quote and the spot exchange rate quote—that is, the forward premium or discount, with the points scaled so that they can be related to the last decimal place in the spot quote. Forward points are adjustments to the spot price of the base currency, using our standard price/base currency notation.

This means that forward rate quotes in professional FX markets are typically shown as the bid/offer on the spot rate and the number of forward points at each maturity.[7] For illustration purposes, let's assume that the bid/offer for the spot and forward points for the USD/EUR exchange rate are as shown in Exhibit 10-1.

[7] *Maturity* is defined in terms of the time between spot settlement—usually T + 2—and the settlement of the forward contract.

EXHIBIT 10-1 Sample Spot and Forward Quotes (Bid/Offer)

Maturity	Spot Rate or Forward Points
Spot (USD/EUR)	1.3549/1.3651
One month	−5.6/−5.1
Three months	−15.9/−15.3
Six months	−37.0/−36.3
Twelve months	−94.3/−91.8

One should note several aspects of this exhibit. First, as always, the offer in the bid/offer quote is larger than the bid. In this example, the forward points are negative (i.e., the forward rate for the EUR is at a discount to the spot rate), but the bid is a smaller number (–5.6, versus −5.1 at the one-month maturity). Second, the absolute number of forward points is an increasing function of the term of the forward contract. Third, because this is an over-the-counter (OTC) market, a client is not restricted to dealing *only* at the dates and maturities shown. Dealers typically quote standard forward dates, but forward deals can be for any forward date the client requires. The forward points for these nonstandard (referred to as "broken") forward dates will typically be interpolated on the basis of the points shown for the standard settlement dates. Fourth, to convert any of these quoted forward points into a forward rate, one would divide the number of points by 10,000 (to scale down to the fourth decimal place, the last decimal place in the USD/EUR spot quote) and then add the result to the spot exchange rate quote.[8] However, one must be careful about which side of the market (bid or offer) is being quoted. For example, suppose a market participant was *selling* the EUR forward against the USD. Given the USD/EUR quoting convention, the EUR is the base currency. This means the market participant must use the *bid* rates (i.e., the market participant will hit the bid) given the USD/EUR quoting convention. Using the data in Exhibit 10-1, the three-month forward *bid* rate in this case would be based on the bid for both the spot and the forward points, and hence would be:

$$1.3549 + \left(\frac{-15.9}{10{,}000}\right) = 1.35331$$

This means that the market participant would be selling EUR three months forward at a price of USD1.35331 per EUR. Fifth, the quoted points are already scaled to each maturity—they are not annualized—so there is no need to adjust them.

The situation is slightly different when calculating forward exchange rates for an FX swap. An FX swap transaction consists of simultaneous spot and forward transactions, where the base currency is being bought spot and sold forward or sold spot and bought forward. FX swaps are used for a variety of purposes, such as swap financing as well as rolling either hedges or speculative positions forward in time as the underlying forward contract matures. Because swaps involve simultaneous and offsetting transactions—one is a buy, the other a sell, in terms of the base currency—a common spot rate is applied to both the spot leg of the transaction

[8]Because the JPY/USD exchange rate is quoted to only two decimal places, forward points for the dollar–yen currency pair are divided by 100.

and to the calculation of the forward rate. Because the client is not being charged a bid/offer spread on the spot rate, it is standard practice to use the midmarket spot exchange rate for the swap transaction. The forward points will still be based on either the bid or the offer, however, depending on whether the market participant is buying or selling the base currency forward. This method of quoting swap pricing is applied whenever a dealer's client transacts simultaneous spot and forward deals in the same base currency.

We now turn to considering what determines the bid/offer spread for forward swap points quoted by dealers to clients. When we discussed *spot* bid/offer rates, we indicated that the bid/offer spread depends on three factors: the interbank market liquidity of the underlying currency pair, the size of the transaction, and the relationship between the client and the dealer. These same factors also apply to bid/offer spreads for forward points. For forward bid/offer spreads, we can also add a fourth factor: the term of the forward contract. Generally, the longer the term of the forward contract, the wider the bid/offer. This relationship holds because, as the term of the contract increases,

- Liquidity in the forward market tends to decline.
- The exposure to counterparty credit risk increases.
- The interest rate risk of the contract increases (forward rates are based on interest rate differentials, and a longer duration equates to higher price sensitivity to movements in interest rates).

Finally, we consider the mark-to-market of forward contracts. As with other financial instruments, the mark-to-market value of forward contracts reflects the profit (or loss) that would be realized from closing out the position at current market prices. To close out a forward position, it must be offset with an equal and opposite forward position using the spot exchange rate and forward points available in the market when the offsetting position is created. When a forward contract is initiated, the forward rate is such that no cash changes hands (i.e., the mark-to-market value of the contract at initiation is zero). From that moment onward, however, the mark-to-market value of the forward contract will change as the spot exchange rate changes and as interest rates change in either of the two currencies.

Let's look at an example. Suppose that a market participant bought GBP10 million for delivery against the AUD in six months at an all-in forward rate of 1.6100 AUD/GBP. (The all-in forward rate is simply the sum of the spot rate and the forward points, appropriately scaled to size.) Three months later, the market participant wants to close out this forward contract. This would require selling GBP10 million three months forward using the AUD/GBP spot exchange rate and forward points in effect at that time.[9] Assume the bid/offer for spot and forward points three months prior to the settlement date are:

Spot rate (AUD/GBP)	1.6210/1.6215
Three-month points	130/140

[9]Note that the offsetting forward contract is defined in terms of the original position taken: The original position in this example was long GBP 10 million, so the offsetting contract is short GBP 10 million. There is an ambiguity here, however: To be *long* GBP 10 million at 1.6100 AUD/GBP is equivalent to being *short* AUD 16,100,000 (10,000,000 × 1.6100) at the same forward rate. To avoid this ambiguity, for the purposes of this chapter we will state what the relevant forward position is for mark-to-market purposes. The net gain or loss from the transaction will be reflected in the alternate currency.

To sell GBP (the base currency in the AUD/GBP quote), we will be calculating the *bid* side of the market. Hence, the appropriate all-in three-month forward rate to use is:

$$1.6210 + 130/10{,}000 = 1.6340$$

This means that the market participant originally bought GBP10 million at an AUD/GBP rate of 1.6100 and subsequently sold them at a rate of 1.6340. These GBP amounts will net to zero at settlement date (GBP10 million both bought and sold), but the AUD amounts will not, because the forward rate has changed. The AUD cash flow at settlement date will equal

$$(1.6340 - 1.6100) \times 10{,}000{,}000 = +\text{AUD}240{,}000$$

This is a cash *inflow* because the market participant was long the GBP with the original forward position and the GBP subsequently appreciated (the AUD/GBP rate increased).

This cash flow will be paid at settlement day, which is still three months away. To calculate the mark-to-market on the dealer's position, this cash flow must be discounted to the present. The present value of this amount is found by discounting the settlement day cash flow by the three-month discount rate. Because this amount is in AUD, we use the three-month AUD discount rate. Let's use LIBOR and suppose that three-month AUD LIBOR is 4.80 percent (annualized). The present value of this future AUD cash flow is then:

$$\frac{\text{AUD}240{,}000}{1 + 0.048\left(\frac{90}{360}\right)} = \text{AUD}237{,}154$$

This is the mark-to-market value of the original long GBP10 million six-month forward when it is closed out three months prior to settlement.

To summarize, the process for marking to market a forward position is relatively straightforward and involves four steps:

1. Create an equal and offsetting forward position to the original forward position. (In the preceding example, the market participant was long GBP 10 million forward, so the offsetting forward contract would be to sell GBP 10 million.)
2. Determine the appropriate all-in forward rate for this new, offsetting forward position. If the base currency of the exchange rate quote is being sold, then use the bid side of the market; if it is being bought, use the offer.
3. Calculate the cash flow at settlement day. This amount will be based on the original contract size times the difference between the original forward rate and that calculated in step 2. If the currency the market participant was originally long (or short) subsequently appreciated (or depreciated), then there will be a cash *inflow*. Otherwise there will be is a cash *outflow*. (In the preceding example, the market participant was long the GBP, which subsequently appreciated, leading to a cash inflow at settlement day.)
4. Calculate the present value of this cash flow at the future settlement date. The currency of the cash flow and the discount rate must match. (In the example, the cash flow at the settlement date was in AUD, so an AUD LIBOR was used to calculate the present value.)

EXAMPLE 10-3 Forward Rates and the Mark-to-Market of Forward Positions

Six months ago, a dealer sold CHF1 million forward against the GBP for a 180-day term at an all-in rate of 1.4850 (CHF/GBP). Today, the dealer wants to roll this position forward for another six months (i.e., the dealer will use an FX swap to roll the position forward). The following are the current spot rate and forward points being quoted for the CHF/GBP currency pair:

Spot rate (CHF/GBP)	1.4939/1.4941
One month	−8.3/−7.9
Two months	−17.4/−16.8
Three months	−25.4/−24.6
Four months	−35.4/−34.2
Five months	−45.9/−44.1
Six months	−56.5/−54.0

1. The current all-in bid rate for delivery of GBP against the CHF in three months is *closest* to:
   A. 1.49136.
   B. 1.49150.
   C. 1.49164.
2. The cash flow that the dealer will realize on the settlement date is *closest* to an:
   A. inflow of GBP4,057.
   B. inflow of GBP8,100.
   C. outflow of GBP5,422.
3. The all-in rate that the dealer will use today to sell the CHF six months forward against the GBP is *closest* to:
   A. 1.48825.
   B. 1.48835.
   C. 1.48860.
4. Some time ago, Laurier Bay Capital, an investment fund based in Los Angeles, hedged a long exposure to the New Zealand dollar by selling NZD10 million forward against the USD; the all-in forward price was 0.7900 (USD/NZD). Three months prior to the settlement date, Laurier Bay wants to mark this forward position to market. The bid/offer for the USD/NZD spot rate, the three-month forward points, and the three-month LIBORs (annualized) are as follows:

Spot rate (USD/NZD)	0.7825/0.7830
Three-month points	−12.1/–10.0
Three-month LIBOR (NZD)	3.31%
Three-month LIBOR (USD)	0.31%

The mark-to-market for Laurier Bay's forward position is *closest* to:
A. −USD87,100.
B. +USD77,437.
C. +USD79,938.

5. Now suppose that instead of having a long exposure to the NZD, Laurier Bay Capital had a long forward exposure to the USD, which it hedged by selling USD10 million forward against the NZD at an all-in forward rate of 0.7900 (USD/NZD). Three months prior to settlement date, it wants to close out this short USD forward position. Using the table in Question 4, the mark-to-market for Laurier Bay's short USD forward position is *closest* to:
A. −NZD141,117.
B. −NZD139,959.
C. −NZD87,100.

Solution to 1: A is correct. The current all-in three month bid rate for GBP (the base currency) is equal to 1.4939 + (–25.4/10,000) = 1.49136.

Solution to 2: A is correct because 180 days ago, the dealer sold 1 million CHF against the GBP for 1.4850. Today, the dealer will have to buy CHF1 million to settle the maturing forward contract, so the CHF amounts will net to zero on settlement day. Because these CHF amounts net to zero, the cash flow on settlement day is measured in GBP. The GBP amount is calculated as follows: 180 days ago, the dealer sold CHF1 million against the GBP at a rate of 1.4850, which is equivalent to buying GBP673,400.67 (1,000,000/1.4850). That is, based on the forward contract, the dealer will receive GBP673,400.67 on settlement day. Today, the dealer is buying CHF1 million at a spot rate of 1.4940 (the midmarket spot rate, because this is an FX swap). This transaction is equivalent to selling GBP669,344.04 (1,000,000/1.4940). That is, based on the spot transaction, the dealer will pay out GBP669,344.04 on settlement day. Combining these two legs of the swap transaction, we have:

$$\frac{1{,}000{,}000}{1.4850} - \frac{1{,}000{,}000}{1.4940} = \text{GBP4,056.63}$$

This is a cash inflow for the dealer because the dealer went short the CHF (long the GBP) and the CHF depreciated against the GBP (equivalently, the GBP appreciated against the CHF) over the life of the forward contract.

Solution to 3: C is correct. The dealer will sell CHF against the GBP, which is equivalent to buying GBP (the base currency) against the CHF. Hence the *offer* side of the market will be used for forward points, and because this is an FX swap, the midmarket on the spot quote will be used. Hence the all-in forward price will be 1.4940 + (–54.0/10,000) = 1.48860.

Solution to 4: C is correct. Laurier Bay sold NZD10 million forward to the settlement date at an all-in forward rate of 0.7900 (USD/NZD). To mark this position to market, it would need an offsetting forward transaction involving buying NZD10 million three months forward to the settlement date. The NZD amounts on settlement date net to zero. For the offsetting forward contract, because the NZD is the base currency in the USD/NZD quote, buying NZD forward means paying the offer for both the spot rate and forward

points. This leads to an all-in three-month forward rate of $0.7830 - 0.0010 = 0.7820$. On settlement day, Laurier Bay will receive USD7,900,000 (NZD10,000,000 × 0.7900 USD/NZD) from the original forward contract and pay out USD7,820,000 (NZD10,000,000 × 0.7820 USD/NZD) based on the offsetting forward contract. This gives a net cash flow on settlement day of $10{,}000{,}000 \times (0.7900 - 0.7820) =$ +USD80,000.

This is a cash inflow because Laurier Bay sold the NZD forward and the NZD depreciated against the USD. This USD cash inflow will occur in three months. To calculate the mark-to-market value of the original forward position, we need to calculate the present value of this USD cash inflow using the three-month USD discount rate (we use USD LIBOR for this purpose):

$$\frac{\text{USD80,000}}{1 + 0.0031\left(\frac{90}{360}\right)} = +\text{USD79,938}$$

Solution to 5: B is correct. This is because Laurier Bay initially sold USD10 million forward, and it will have to buy USD10 million forward to the same settlement date (i.e., in three months' time) in order to close out the initial position. Buying USD using the USD/NZD currency pair is the same as selling the NZD. Because the NZD is the base currency in the USD/NZD quote, selling the NZD means calculating the *bid* rate:

$$0.7825 + (-12.1/10{,}000) = 0.78129$$

At settlement, the USD amounts will net to zero (10 million USD both bought and sold). The NZD amounts will not net to zero, however, because the all-in forward rate changed between the time Laurier Bay initiated the original position and when it closed out this position. At initiation, Laurier Bay contracted to sell USD10,000,000 and receive NZD12,658,228 (i.e., 10,000,000/0.7900) on the settlement date. To close out the original forward contract, Laurier Bay entered into an offsetting forward contract to receive USD10,000,000 and pay out NZD12,799,345 (i.e., 10,000,000/0.78129) at settlement. The difference between the NZD amounts that Laurier Bay will receive and pay out on the settlement date equals:

$$\text{NZD12,658,228} - \text{NZD12,799,345} = -\text{NZD141,117}$$

This is a cash *outflow* for Laurier Bay because the fund was *short* the USD in the original forward position and the USD subsequently *appreciated* (i.e., the NZD subsequently depreciated, because the all-in forward rate in USD/NZD dropped from 0.7900 to 0.78129). This NZD cash outflow occurs in three months' time, and we must calculate its present value using the three-month NZD LIBOR:

$$\frac{-\text{NZD141,117}}{1 + 0.0331\left(\frac{90}{360}\right)} = -\text{NZD139,959}$$

3. A LONG-TERM FRAMEWORK FOR EXCHANGE RATES

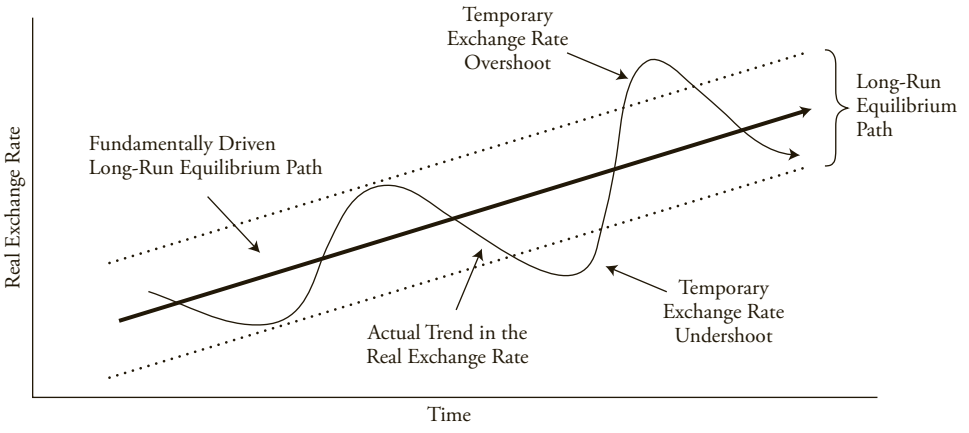
Having reviewed the basic tools of the FX market, we now turn our focus to how they are used in practice. At the heart of the trading decision in FX (and other) markets lies a view on future market prices and conditions. This outlook guides the market participant's decisions with respect to risk exposures, as well as whether currency hedges should be implemented and, if so, how they should be managed. Even the decision to be fully hedged, or to have no market exposure at all, implies an opinion that future market conditions are so uncertain that no sufficiently profitable active trading opportunities are available.

In this and the following sections, we lay out a framework for developing a view about future exchange rate movements, which should serve as a guide for how FX positions should be managed. We begin by examining international parity conditions, which describe the interrelationships that jointly determine *long-run* movements in exchange rates, interest rates, and inflation. These parity conditions are the basic building blocks for describing long-term equilibrium levels for exchange rates. In subsequent sections, we expand beyond this simple view of how exchange rates are determined in a long-term, ideal world by bringing in a broader view of real-world factors that must be considered in forming an intelligent market opinion, especially over medium- and short-term horizons.

Always keep in mind that exchange rate movements reflect complex interactions among multiple forces. In trying to untangle this complex web of interactions, we must clearly delineate the following four concepts:

1. *Long run versus short run.* Many of the factors that determine exchange rate movements exert subtle but persistent influences over long periods of time. Although a poor guide for short-term prediction, longer-term equilibrium values nevertheless act as an anchor for exchange rate movements. Exhibit 10-2 shows a stylized representation of movements in spot exchange rates within a convergence path centered on a long-run equilibrium

EXHIBIT 10-2 A Stylized Model of the Long-Term Trend in a Currency's Value



Source: Rosenberg (2002, 32).

exchange rate. The forces that affect long-run convergence to fair value are discussed more in the sections on international parity conditions and long-run equilibrium models. Subsequent sections discuss factors that help determine the medium- and short-term deviations within this convergence channel.

2. *Real versus nominal values.* Observable exchange rates (and interest rates) represent the prices of tradable financial products, but they are not inflation adjusted. In contrast, people save and invest to facilitate future purchases of *real* (inflation-adjusted) goods and services. Therefore, movements in real variables exert considerable influence over nominal variables, including nominal spot exchange rates. Hence, in some of the subsequent material the focus will be on real variables in order to better conceptualize exchange rate determinants.

   Nonetheless, only *nominal* exchange rates are tradable or used in mark-to-market calculations. Hence, it is necessary to be able to map expected movements in real rates back into expected future movements in nominal exchange (and interest) rates. Being able to form an opinion about future nominal exchange rate movements is the ultimate goal of the material covered in this and all subsequent sections of this chapter.
3. *Expected versus unexpected changes.* In reasonably efficient markets, prices will adjust to reflect market participants' expectations of future developments. When a key factor, say inflation, is trending gradually in a particular direction, market pricing will eventually come to reflect expectations that this trend will continue. In contrast, large, unexpected movements in a variable (for example, a central bank intervening in the foreign exchange market) can lead to immediate, discrete price adjustments. This concept of expected versus unexpected changes is closely related to what might broadly be referred to as risk. For example, a moderate but steady rate of inflation will not have the same effect on market participants as an inflation rate that is very unpredictable. The latter clearly describes a riskier financial environment. Market pricing will reflect risk premiums—that is, the compensation that traders and investors demand for being exposed to unpredictable outcomes. Whereas expectations of long-run equilibrium values tend to evolve slowly, risk premiums—which are closely related to confidence and reputation—can change quickly in response to unexpected developments.
4. *Relative movements.* An exchange rate represents the relative price of one currency in terms of another. Hence, for exchange rate determination, the levels or variability of key factors in any particular country are typically much less important than the *differences* in these factors across countries. For example, knowing that inflation is increasing in Country A may not give much insight into the direction of the A/B exchange rate without also knowing what is happening with the inflation rate in Country B.

As a final word of caution—and this cannot be emphasized enough: *There is no simple formula, model, or approach that will allow market participants to precisely forecast exchange rates* (or any other financial prices) or to be able to make all trading decisions profitable. We live in an imperfect world where exchange rates and other financial prices can be highly erratic and hard to explain even after the fact, let alone predict in advance. Nonetheless, to operate in financial markets requires acceptance of these imperfections. It also requires that market participants have a market view to guide their decisions, even if this view requires significant revision as new information becomes available. The following sections provide a framework for formulating a view on FX markets, a guide for thinking through the complex forces driving exchange rates. As with all theory, however, it does not eliminate the need for insightful analysis of actual economic and market conditions.

3.1. International Parity Conditions

International parity conditions form the building blocks of most models of exchange rate determination. There are seven key international parity conditions:

1. Covered interest rate parity.
2. Uncovered interest rate parity.
3. Forward exchange rates as unbiased predictors of future spot exchange rates.
4. Purchasing power parity.
5. The Fisher effect.
6. The international Fisher effect.
7. Real interest rate parity.

Parity conditions show how expected inflation differentials, interest rate differentials, forward exchange rates, current spot exchange rates, and expected future spot exchange rates would be linked in an ideal world. These conditions typically make simplifying assumptions, such as perfect information that is available to all market participants, risk neutrality, and freely adjustable market prices. If these international parity conditions held at all times, moving capital from one market to another would offer no profitable trading opportunities for international investors.

Most empirical studies find, however, that the key international parity conditions rarely hold in either the short or the medium term. The exception is covered interest rate parity, which is the only one of the parity conditions that is enforced by an executable arbitrage relationship. There are often significant and persistent departures from purchasing power parity, uncovered interest rate parity, and real interest rate parity. In addition, forward exchange rates have typically been found to be poor predictors of future spot exchange rates.

The failure of international parity conditions raises an interesting question: Why bother to study them at all if they do not work? There are essentially two reasons. First, although the conditions are frequently violated, each reflects economic forces that should not be ignored altogether. Second, and perhaps even more importantly, from a trader/investor/analyst's perspective, international parity conditions truly become interesting only when they fail to hold, because it is only then that currency positions offer profitable opportunities. As mentioned earlier, the parity relationships are one of the most basic frameworks from which a more broadly based, long-term view of future market conditions can be constructed. Such a market outlook is a prerequisite for a market participant to manage longer-term risk exposures.

3.1.1. Covered Interest Rate Parity

We have already discussed covered interest rate parity in our examination of forward exchange rates. This parity condition describes a riskless arbitrage relationship in which *an investment in a foreign money market instrument that is completely hedged against exchange rate risk should yield exactly the same return as an otherwise identical domestic money market investment*. Given the spot exchange rate and the domestic and foreign yields, the forward exchange rate must equal the rate that gives these two alternative investment strategies—invest either in a domestic money market instrument or in a fully currency-hedged foreign money market instrument—exactly the same holding period return.[10] If one strategy gave a higher holding period return

[10]Note that the spot exchange rate, the forward exchange rate, and the two interest rates are actually determined simultaneously in the market. One should not think of one of them, say the forward exchange rate, being determined by the others.

than the other, then an investor could short-sell the lower-yielding approach and invest the proceeds in the higher-yielding approach, earning riskless profits in the process. In real-world financial markets, such a disparity would be quickly arbitraged away by alert market participants.

For covered interest rate parity to hold exactly, it must be assumed that there are zero transaction costs and that the underlying domestic and foreign money market instruments being compared are identical in terms of liquidity, maturity, and default risk. In most cases where capital is permitted to flow freely, spot and forward exchange markets are liquid, and financial market conditions are relatively stress-free, covered interest rate differentials are generally found to be close to zero; that is, Equation 10-1 in Section 2.2 holds.

3.1.2. Uncovered Interest Rate Parity

According to the **uncovered interest rate parity** condition, the *expected* return on an uncovered (i.e., unhedged) foreign currency investment should equal the return on a comparable domestic currency investment. The return on a risk-free domestic money market instrument is known with certainty for a domestic investor: It is the money market instrument's yield. However, an unhedged investment in a foreign currency money market instrument exposes a domestic investor to the risk of spot exchange rate movements between the time the investment is made and when it matures. Uncovered interest rate parity states that *the change in spot rate over the investment horizon should, on average, equal the differential in interest rates between the two countries. That is, the expected appreciation or depreciation of the exchange rate just offsets the yield differential, implying that the current forward exchange rate is an unbiased (i.e., correct on average) predictor of the future spot rate*.

An example will help illustrate how uncovered interest rate parity works. To explain the intuition behind this concept more easily, let's switch, as we did with the examples for covered interest parity, from the standard price/base currency notation (P/B) to foreign/domestic currency notation (*f/d*) in order to emphasize the choice between foreign and domestic investments. As before, we also will assume that for the investor the base and domestic currencies are the same.

For our example, assume that this investor has a choice between owning a one-year domestic money market instrument and a one-year foreign-currency-denominated money market investment. We assume that this investor will not hedge the FX risk in the forward exchange market. Under the assumption of uncovered interest rate parity, the investor must compare the *known* return on the domestic investment with the *expected* all-in return on the unhedged foreign-currency-denominated investment (i.e., including the foreign yield as well as movements in the exchange rate, in $S_{f/d}$ terms). The ultimate investment choice between these two investments (domestic and foreign) will depend on which market offers the higher expected return on an unhedged basis.

To be concrete, let's assume that the return on the one-year foreign money market instrument is 10 percent, while the return on the one-year domestic money market instrument is 5 percent. From the investor's perspective, the expected return on the one-year domestic investment in domestic currency terms is known with complete certainty (i.e., 5 percent). This is not the case for the uncovered investment in the foreign currency money market instrument. In domestic currency terms, the investment return on an uncovered (or unhedged) foreign-currency-denominated investment is approximately equal to:[11]

[11]The symbol $\cong$ indicates approximate equality. The approximation holds because the product ($i \times \%\Delta S$) is small compared to the interest rate (i) and the percentage change in the exchange rate ($\%\Delta S$). For simplicity of exposition, we will use the $\cong$ symbol when we introduce an approximation but will subsequently treat the relationship as an equality ($=$) unless the distinction is important for the issue being discussed.

$$(1 + i_f)(1 - \%\Delta S_{f/d}) - 1 \cong i_f - \%\Delta S_{f/d}$$

The percentage change in $S_{f/d}$ enters with a minus sign because an *increase* in $S_{f/d}$ means the foreign currency *declines* in value, reducing the all-in return from the domestic currency perspective of our investor. An increase in $S_{f/d}$ means the base currency, the domestic currency in this case, is appreciating, and that the price (foreign) currency is depreciating.

This all-in return depends on *future* movements in the $S_{f/d}$ rate, which cannot be known until the end of the period. Consider three cases:

1. The $S_{f/d}$ rate is expected to remain unchanged.
2. The domestic currency is expected to appreciate by 10 percent.
3. The domestic currency is expected to appreciate by 5 percent.

In the first case, the investor would prefer the foreign-currency-denominated money market investment because it offers a 10 percent (= 10% – 0%) expected return, while the comparable domestic investment offers only 5 percent. In the second case, the investor would prefer the domestic investment because the expected return on the foreign-currency-denominated investment is 0 percent (= 10% – 10%). In the third case, uncovered interest rate parity holds because both investments offer a 5 percent (= 10% – 5%) expected return. In this case, the investor is assumed to be indifferent between the alternatives.

Note that in the third case, in which uncovered interest rate parity holds, although the *expected* return over the one-year investment horizon is the same, the *distribution* of possible total return outcomes is quite different. For our investor, the return on the domestic money market instrument is known with certainty. In contrast, the distribution of the all-in return on the foreign money market instrument embodies uncertainty with respect to the future $S_{f/d}$ rate. Hence, when we say that the investor would be indifferent between owning domestic and foreign investments because they both offer the same *expected* return (5 percent), we are assuming that the investor is *risk neutral*. Thus, uncovered interest rate parity assumes that there are enough risk-neutral investors to force equality of expected returns.

In terms of our example's foreign/domestic (*f/d*) notation, uncovered interest rate parity asserts

$$i_f - \%\Delta S^e_{f/d} = i_d$$

where $\%\Delta S^e_{f/d}$ is the expected change in the foreign currency price of the domestic currency over the investment horizon. This equation can be rearranged to restate the uncovered interest rate parity condition in terms of the expected change in the exchange rate:[12]

$$\%\Delta S^e_{f/d} = i_f - i_d \tag{10-2}$$

According to this equation, the expected change in the spot exchange rate over the investment horizon should be reflected in the interest rate differential. Using our example, if the yield spread between the foreign and domestic investments is 5 percent ($i_f - i_d = 5\%$),

[12]The mathematically inclined reader may note that this equation cannot hold simultaneously for $S_{f/d}$ and $S_{d/f}$ (= $1/S_{f/d}$) because their percentage changes are not of exactly equal magnitude. This reflects our earlier approximation. Using the exact return on the unhedged foreign instrument would alleviate this issue at the expense of a less intuitive equation.

then this spread implicitly reflects the expectation that the domestic currency will strengthen versus the foreign currency by 5 percent ($\%\Delta S^{e}_{f/d} = 5\%$).

Note that with uncovered interest rate parity, the country with the *higher* interest rate or money market yield is expected to see the value of its currency *depreciate*. It is this depreciation of the currency that offsets the higher yield so that the (expected) all-in return on the two investment choices is the same. Hence, if the uncovered interest rate parity condition held consistently in the real world, it would rule out the possibility of earning excess returns from going long a high-yield currency and going short a low-yield currency. If uncovered interest rate parity held, the depreciation of the high-yield currency would exactly offset the yield advantage that the high-yield currency offers. Taking this to its logical conclusion, if uncovered interest rate parity held at all times, investors would have no incentive to shift capital from one currency to another because expected returns on otherwise identical money market investments would be equal across markets and risk-neutral investors would be indifferent among them.

Most studies find that uncovered interest rate parity fails to hold over short- and medium-term periods, although there is evidence that it works better over very long-term horizons. Over short- and medium-term periods, interest rate differentials have generally been found to be poor predictors of future exchange rate changes. Indeed, most studies find that high-yield currencies fail to weaken in line with the path predicted by uncovered interest rate parity. That is, the rate of depreciation of the high-yield currency has been found to be less than the implied uncovered interest rate parity path. In many cases, high-yield currencies have been found to *strengthen*, not weaken, conflicting even more strongly with the implication of uncovered interest rate parity.

Such findings have significant implications for foreign exchange investment strategies. If high-yield currencies fail to depreciate in line with the path predicted by the uncovered interest rate parity condition, then high-yield currencies should exhibit a tendency to outperform low-yield currencies over time. If so, investors could adopt strategies that overweight high-yield currencies at the expense of low-yield currencies and generate attractive returns in the process. Such approaches are known as FX carry trade strategies. We discuss them in greater depth in Section 4.

3.1.3. Spot and Forward Rates as Predictors of Future Spot Rates

The covered interest rate parity condition describes the relationship among the spot exchange rate, the forward exchange rate, and interest rates. Let's keep using the foreign/domestic exchange rate notation (*f/d*) to simplify the explanation. As we illustrated in Section 2, the arbitrage condition that underlies covered interest rate parity can be rearranged to give an expression for the forward premium or discount:

$$F_{f/d} - S_{f/d} = S_{f/d}\left[\frac{\left(\frac{Actual}{360}\right)}{1 + i_d\left(\frac{Actual}{360}\right)}\right](i_f - i_d)$$

The domestic currency will trade at a forward premium ($F_{f/d} > S_{f/d}$) if, and only if, the foreign risk-free interest rate exceeds the domestic risk-free interest rate ($i_f > i_d$).

For the sake of simplicity, assume that the investment horizon is one year, leading to:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{i_f - i_d}{1 + i_d}\right) \cong S_{f/d}(i_f - i_d)$$

This equation can be reexpressed to show the forward discount or premium as a percentage of the spot rate:

$$\frac{F_{f/d} - S_{f/d}}{S_{f/d}} \cong i_f - i_d$$

As we showed previously, if uncovered interest parity holds—that is, investors are risk neutral—then the expected change in the spot rate is equal to the interest rate differential:

$$\%\Delta S^e_{f/d} = i_f - i_d$$

We can link these two equations by assuming that uncovered interest rate parity holds. If this is the case, then the forward premium (discount) on a currency, expressed in percentage terms, equals the expected percentage appreciation or depreciation of the domestic currency:

$$\frac{F_{f/d} - S_{f/d}}{S_{f/d}} = \%\Delta S^e_{f/d} = i_f - i_d$$

It follows that the forward exchange rate equals the expected future spot exchange rate:

$$F_{f/d} = S^e_{f/d}$$

Thus, in theory, *the forward exchange rate will be an unbiased forecast of the future spot exchange rate if both covered and uncovered interest rate parity hold.*

In our previous example, foreign interest rates were assumed to be 5 percent higher (10% – 5%) than domestic interest rates. *Uncovered* interest parity would imply that the domestic currency was expected to appreciate by 5 percent against the foreign currency (i.e., $\%\Delta S^e_{f/d} = 5\%$). *Covered* interest rate parity would imply that the domestic currency must trade at a 5 percent forward premium; that is, $(F_{f/d} - S_{f/d}) = 5\%$. The latter must hold because it is enforced by arbitrage. *So, asking whether the forward exchange rate is an unbiased predictor of the spot exchange rate is the same as asking whether uncovered interest rate parity holds*.

How might uncovered interest rate parity, and with it equality of the forward exchange rate and the expected future spot exchange rate, be enforced? It is not enforced by arbitrage, because there is no combination of trades that will lock in a certain profit. If the forward rate is above (or below) speculators' expectation of the future spot rate, however, then risk-neutral speculators will buy the domestic currency in the spot (or forward) market and simultaneously sell it in the forward (or spot) market. If their expectations are correct, they will make a profit, on average. These transactions will push the forward premium into alignment with the consensus expectation of the future spot rate.

Note, however, that spot exchange rates are volatile and determined by a complex web of influences—interest rate differentials are only one among many factors. In other words, speculators can also lose. Moreover, speculators are rarely, if ever, truly risk neutral, and without an arbitrage relationship to enforce it, uncovered interest rate parity is often violated. *In general, this means that forward exchange rates are poor predictors of future spot exchange rates*.

Current spot exchange rates are not very good predictors of future spot exchange rates, either. Superficially, it might seem that spot rates could be used as predictors, but movements in spot rates are so volatile that they often approximate a random walk:

$$S_{t+1} - S_t = \Delta S_{t+1} \cong \varepsilon_{t+1}$$

If the distribution for the error term has a mean of zero, then $E_t(\varepsilon_{t+1}) = 0$. This would imply that the expectation for the future spot exchange rate would be the current spot exchange rate $[E_t(S_{t+1}) = S_t]$. In practice, however, current spot rates are poor predictors of future spot rates because of the high volatility in exchange rate movements: Future spot exchange rates rarely equal current spot exchange rates and are often not even close. Also, without using *any* current information at all to attempt to predict future spot rates (such as current interest rate differentials), the random walk prediction can be slightly biased, on average.

EXAMPLE 10-4 Covered and Uncovered Interest Parity, Predictors of Future Spot Rates

An Australian-based fixed income asset manager is deciding whether to allocate money between Australia and Japan. Note that the base currency in the exchange rate quote (AUD) is the domestic currency for the asset manager.

JPY/AUD spot rate (midmarket)	79.25
One-year forward points (midmarket)	−301.9
One-year Australian deposit rate	5.00%
One-year Japanese deposit rate	1.00%

1. Based on uncovered interest rate parity, over the next year, the expected change in the JPY/AUD rate is *closest* to a(n):
   A. decrease of 10 percent.
   B. decrease of 4 percent.
   C. increase of 4 percent.
2. The *best* explanation of why this prediction may not be very accurate is that:
   A. covered interest parity does hold in this case.
   B. the forward points indicate that a riskless arbitrage opportunity exists.
   C. there is no arbitrage condition that forces uncovered interest rate parity to hold.
3. Using the forward points to forecast the future JPY/AUD spot rate one year ahead assumes that:
   A. investors are risk neutral.
   B. spot rates follow a random walk.
   C. it is not necessary for uncovered interest rate parity to hold.
4. Forecasting that the JPY/AUD spot rate one year from now will equal 79.25 assumes that:
   A. investors are risk neutral.
   B. spot rates follow a random walk.
   C. it is necessary for uncovered interest rate parity to hold.

5. If the asset manager completely hedged the currency risk associated with a one-year Japanese deposit using a forward rate contract, the one-year all-in holding return, in AUD, would be *closest* to:
   A. 0 percent.
   B. 1 percent.
   C. 5 percent.

The fixed income manager collects the following information, and uses it along with the international parity conditions in order to estimate investment returns and future exchange rate movements.

Today's One-Year LIBOR		Currency Pair	Spot Rate Today
JPY	0.10%	JPY/USD	81.30
USD	0.10%	USD/GBP	1.5950
GBP	3.00%	JPY/GBP	129.67

6. If covered interest parity holds, the all-in, one-year investment return to a Japanese investor whose currency exposure to the GBP is fully hedged is *closest* to:
   A. 0.10 percent.
   B. 0.17 percent.
   C. 3.00 percent.
7. If uncovered interest parity holds, today's expected value for the JPY/GBP currency pair one year from now would be *closest* to:
   A. 126.02.
   B. 129.67.
   C. 130.05.
8. If uncovered interest parity holds, between today and one year from now the expected movement in the JPY/USD currency pair is *closest* to:
   A. −1.60 percent.
   B. +0.00 percent.
   C. +1.63 percent.

Solution to 1: B is correct. The expected depreciation of the Australian dollar (decline in the JPY/AUD rate) is equal to the interest rate differential between Australia and Japan (5% – 1%).

Solution to 2: C is correct. There is no arbitrage condition that forces uncovered interest rate parity to hold. In contrast, arbitrage virtually always ensures that covered interest rate parity holds. This is the case for our table, where the −302-point discount is calculated from the covered interest rate parity equation.

Solution to 3: A is correct. Using forward rates (i.e., adding the forward points to the spot rate) to forecast future spot rates assumes that uncovered interest rate parity holds. In turn, uncovered interest rate parity assumes that investors are risk neutral. If these conditions hold, then movements in the spot exchange rate, although they *approximate*

a random walk, will not actually be a random walk because current interest spreads will determine expected exchange rate movements.

Solution to 4: B is correct. Assuming that the current spot exchange rate is the best predictor of future spot rates assumes that exchange rate movements follow a random walk. If uncovered interest rate parity holds, the current exchange rate will not be the best predictor unless the interest rate differential happens to be zero. Risk neutrality is needed to enforce uncovered interest rate parity, but it will not make the current spot exchange rate the best predictor of future spot rates.

Solution to 5: C is correct. A fully hedged JPY investment would provide the same return as the AUD investment: 5 percent. This is covered interest rate parity, an arbitrage condition.

Solution to 6: A is correct. If covered interest rate parity holds (and it very likely does, because this is a pure arbitrage relationship), then the all-in investment return to a Japanese investor in a one-year, fully hedged GBP LIBOR position would be identical to a one-year JPY LIBOR position: 0.10 percent. No calculations are necessary.

Solution to 7: A is correct. If uncovered interest rate parity holds, then the expected spot rate one year forward is equal to the one-year forward exchange rate. This forward rate is calculated in the usual manner, given the spot exchange rates and LIBORs:

$$S^e = F = 129.67\left(\frac{1.001}{1.03}\right) = 126.02$$

Solution to 8: B is correct. Given uncovered interest rate parity, the expected change in a spot exchange rate is equal to the interest rate differential. At the one-year term, there is no difference between USD and JPY LIBOR.

3.1.4. Purchasing Power Parity

So far, we have looked at the interrelationships between exchange rates and interest rate differentials. Now we turn to examining the relationship between exchange rates and inflation differentials. The basis for this relationship is known as purchasing power parity (PPP).

Various versions of PPP exist. The foundation for all of the versions is the law of one price. According to the **law of one price**, identical goods should trade at the same price across countries when valued in terms of a common currency. To simplify the explanation, as we did with our examples for covered and uncovered interest rate parity, let's continue to use the foreign/domestic currency quote convention (*f/d*) and the case where the base currency in the P/B notation is the domestic currency for the investor in the *f/d* notation.

The law of one price asserts that the foreign price of good $x(P_f^x)$ should equal the exchange-rate-adjusted price of the identical good in the domestic country (P_d^x):

$$P_f^x = S_{f/d} \times P_d^x$$

For example, for a EUR-based consumer, if the price of good x in the euro area is €100 and the nominal exchange rate stands at 1.40 USD/EUR, then the price of good x in the United States should equal $140.

The **absolute version of PPP** simply extends the law of one price to the broad range of goods and services that are consumed in different countries. Using our previous example but now expanded to include all goods and services, not just good x, the broad price level of the foreign country (P_f) should equal the currency-adjusted broad price level in the domestic country (P_d):

$$P_f = S_{f/d} \times P_d$$

This equation implicitly assumes that all domestic and foreign goods are tradable and that the domestic and foreign price indices include the same bundle of goods and services with the same exact weights in each country. Rearranging this equation and solving for the nominal exchange rate ($S_{f/d}$), the absolute version of PPP states that the nominal exchange rate will be determined by the ratio of the foreign and domestic broad price indexes:

$$S_{f/d} = P_f / P_d$$

The absolute version of PPP asserts that the equilibrium exchange rate between two countries is determined entirely by the ratio of their national price levels. However, it is highly unlikely that one will find that this relationship actually holds in the real world. The absolute version of PPP assumes that goods arbitrage will equate the prices of all goods and service across countries, but if transaction costs are significant and/or not all goods and services are tradable, then goods arbitrage will be incomplete. Hence, sizable and persistent departures from absolute PPP are likely.

However, if it is assumed that transaction costs and other trade impediments are constant over time, it might be possible to show that *changes* in exchange rates and *changes* in national price levels are related, even if the relationship between exchange rate *levels* and national price *levels* does not hold. According to the **relative version of PPP**, the percentage change in the spot exchange rate ($\%\Delta S_{f/d}$) will be completely determined by the difference between the foreign and domestic inflation rates ($\pi_f - \pi_d$):[13]

$$\%\Delta S_{f/d} \cong \pi_f - \pi_d \tag{10-3}$$

Intuitively, the relative version of PPP implies that the exchange rate changes to offset changes in competiveness arising from inflation differentials. For example, if the foreign inflation rate is assumed to be 10 percent while the domestic inflation rate is assumed to be 5 percent, then the $S_{f/d}$ exchange rate must rise by 5 percent ($\Delta S_{f/d} = 10\% - 5\% = 5\%$) in order to maintain the relative competitiveness of the two regions. (Stated in an equivalent manner, the currency of the high-inflation country should depreciate relative to the currency of the low-inflation country. Recall that an increase in $\Delta S_{f/d}$ means the domestic currency is appreciating and the foreign currency is depreciating.) If the $S_{f/d}$ exchange rate remained unchanged, the relatively higher foreign inflation rate would erode the competitiveness of foreign companies relative to domestic companies.

[13]We will occasionally need to convert from a relationship expressed in levels of the relevant variables into a relationship among rates of change. If $X = (Y \times Z)$, then $(1 + \%\Delta X) = (1 + \%\Delta Y)(1 + \%\Delta Z)$ and $\%\Delta X \approx \%\Delta Y + \%\Delta Z$, because $(\%\Delta Y \times \%\Delta Z)$ is "small." Similarly, it can be shown that if $X = (Y/Z)$, then $(1 + \%\Delta X) = (1 + \%\Delta Y)/(1 + \%\Delta Z)$ and $\%\Delta X \approx \%\Delta Y - \%\Delta Z$. Applying this to the equation for absolute PPP gives Equation 10-3.

The ***ex ante* version of PPP** follows directly from the relative version of PPP. Whereas the relative version of PPP focuses on *actual* changes in exchange rates being driven by *actual* differences in national inflation rates in a given time period, the *ex ante* version of PPP focuses on *expected* changes in the spot exchange rate being entirely driven by *expected* differences in national inflation rates. *Ex ante* PPP tells us that countries that are expected to run *persistently* high inflation rates should expect to see their currencies depreciate over time, while countries that are expected to run relatively low inflation rates on a sustainable basis should expect to see their currencies appreciate over time. *Ex ante* PPP can be expressed as:

$$\%\Delta S^{e}_{f/d} = \pi^{e}_{f} - \pi^{e}_{d} \tag{10-4}$$

where it is understood that the use of expectations (the "*e*" superscript) indicates that we are now focused on *future* periods. That is, $\%\Delta S^{e}_{f/d}$ represents the expected percentage change in the spot exchange rate, while π^{e}_{d} and π^{e}_{f} represent the domestic and foreign inflation rates expected to prevail over the same period.

The idea that inflation differentials across countries will cause nominal exchange rates to adjust in order to re-equilibrate real purchasing power is closely related to the concept of real exchange rates. As was covered in the preceding chapters, the **real exchange rate** of a currency captures the real purchasing power of that currency, defined in terms of the amount of *real* goods and services that it can purchase internationally. To derive the real exchange rate, the nominal exchange rate is adjusted for the inflation rate differential between the two countries involved in that currency pair.

An equivalent way of viewing the real exchange rate is that it represents the relative price levels in the domestic and foreign countries expressed in the same currency. Mathematically, we can represent the domestic price level in terms of the foreign currency as:

Domestic price level in foreign currency $= S_{f/d} \times P_d$

As before, the foreign price level expressed in terms of the foreign currency is P_f. The ratio between the domestic and foreign price levels is the real exchange rate $q_{f/d}$:

$$q_{f/d} = \left(\frac{S_{f/d} P_d}{P_f}\right) = S_{f/d}\left(\frac{P_d}{P_f}\right) \tag{10-5}$$

For example, let's examine the case of a domestic consumer wanting to buy foreign goods. This means that the real exchange rate ($q_{f/d}$) will be an increasing function of the nominal spot exchange rate ($S_{f/d}$) and the domestic price level and a decreasing function of the foreign price level. This is written as:

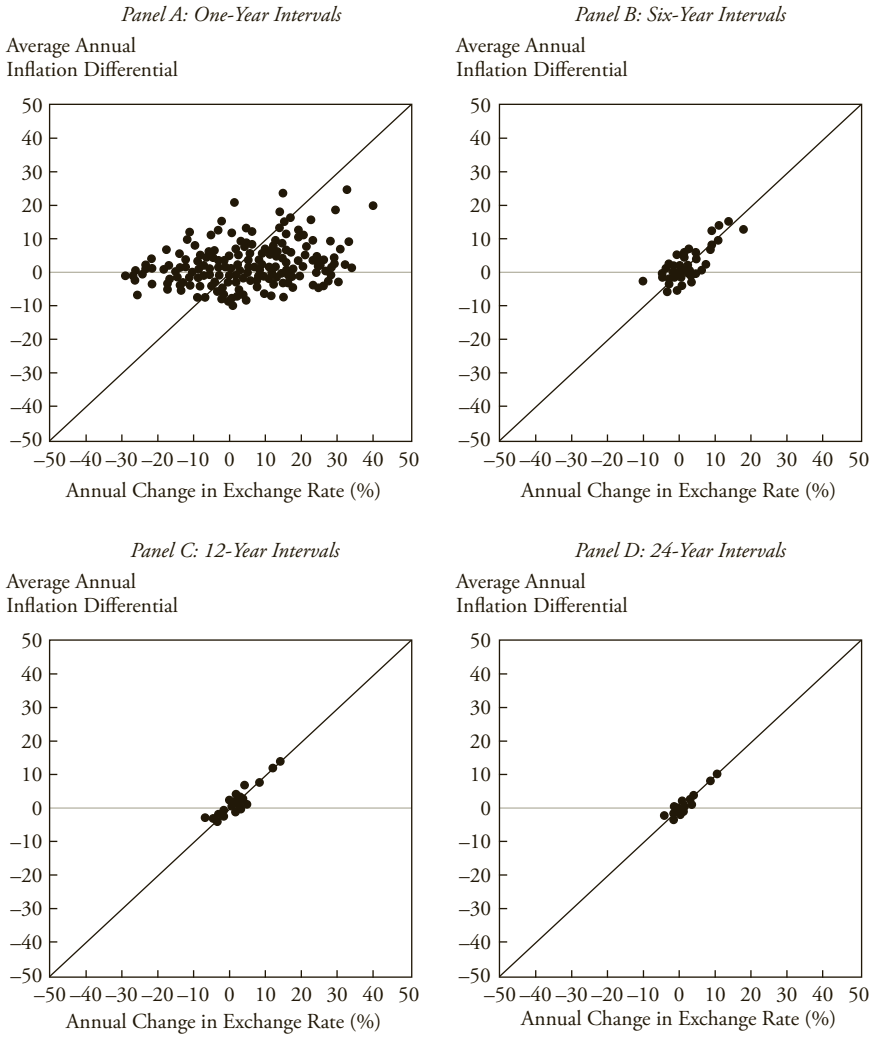
$$q_{f/d} = S_{f/d} \times \left(\frac{CPI_d}{CPI_f}\right)$$

Note that for the domestic consumer, an increasing real exchange rate ($q_{f/d}$) means that it is becoming less expensive, in real terms, to shop in the foreign country: The amount of real goods and services that a domestic consumer can purchase is increasing. Just as an increase in $S_{f/d}$ is an appreciation of the domestic currency in nominal terms, an increase in the real exchange rate $q_{f/d}$ is an appreciation of the domestic currency in real, or inflation-adjusted, terms. (Recall that the domestic currency in our example is the base currency in the standard P/B notation.)

Real exchange rates exhibit a very important empirical property. Studies find that in the *long run*, real exchange rates between countries tend to stabilize around their average values—that is, they mean revert. Stated another way, nominal exchange rates gradually gravitate toward their PPP-based values. This means that, *although over shorter horizons nominal exchange rate movements may appear haphazard, over longer time horizons nominal exchange rates will tend to gravitate toward their long-run PPP equilibrium values.*

Exhibit 10-3 illustrates the success or lack thereof of the relative version of PPP at different time horizons: one year, six years, 12 years, and 24 years. Each chart plots the inflation differential (vertical axis) against the percentage change in the exchange rate (horizontal axis). If PPP holds, the points should fall along the upward-sloping diagonal line. As indicated in the

EXHIBIT 10-3 Impact of Relative Inflation Rates on Exchange Rates over Different Time Horizons



Source: Isard, Faruqee, Kincaid, and Fetherston (2001).

chart, there is no clear relationship between changes in exchange rates and inflation differentials over a one-year time horizon. If the horizon is lengthened to six years and beyond, however, a strong positive relationship becomes apparent. Hence, *PPP appears to be a valid framework to assess long-run fair value in the FX markets, even though the path to PPP equilibrium is excruciatingly slow*. Estimates place the half-life of PPP deviations at around three to five years. That is, on average it takes roughly three to five years to narrow a given PPP deviation by roughly 50 percent.

3.1.5. The Fisher Effect and Real Interest Rate Parity

So far, we have examined the relationship between exchange rates and interest rate differentials between countries, and between exchange rates and inflation differentials. Now we will begin to bring these concepts together by examining how exchange rates, interest rates, and inflation rates interact.

According to what economists call the Fisher effect, one can break down the nominal interest rate (i) in a given country into two parts: (1) the real interest rate in that particular country (r) and (2) the expected inflation rate (π^e) in that country:

$$i = r + \pi^e$$

To relate this concept to exchange rates—the relative price of a currency between *two* countries—we can write the Fisher equation for both the domestic country and a foreign country. If the Fisher effect holds, the nominal interest rates in both countries will equal the sum of their respective real interest rates and expected inflation rates:

$$i_d = r_d + \pi_d^e;$$
$$i_f = r_f + \pi_f^e$$

Because nominal interest rate differentials play an important role in both covered and uncovered interest rate parity calculations, let's take a closer look at the macroeconomic forces that drive the trend in nominal yield spreads. Subtracting the top equation from the bottom equation shows that the nominal yield spread between the foreign and domestic countries ($i_f - i_d$) equals the sum of two parts: (1) the foreign–domestic real yield spread ($r_f - r_d$) and (2) the foreign–domestic expected inflation differential ($\pi_f^e - \pi_d^e$):

$$i_f - i_d = (r_f - r_d) + (\pi_f^e - \pi_d^e)$$

We can rearrange this equation to solve for the foreign–domestic *real* interest rate differential instead of the domestic–foreign *nominal* interest rate differential:

$$(r_f - r_d) = (i_f - i_d) - (\pi_f^e - \pi_d^e)$$

To tie this material to our previous work on exchange rates, let's continue with our previous, simplifying convention of quoting the currencies using foreign/domestic notation (*f/d*). Now, if uncovered interest rate parity holds, then the nominal interest rate spread ($i_f - i_d$) equals the expected change in the exchange rate ($\%\Delta S^e_{f/d}$). Similarly, if *ex ante* PPP holds, the difference in expected inflation rates ($\pi_f^e - \pi_d^e$) also equals the expected change in the exchange rate. Assuming that both uncovered interest rate parity and *ex ante* PPP hold leads to:

$$(r_f - r_d) = \%\Delta S^e_{f/d} - \%\Delta S^e_{f/d} = 0$$

According to this equation, if both uncovered interest rate parity and *ex ante* PPP hold, then the real yield spread between the domestic and foreign countries ($r_f - r_d$) will be zero. If that is the case, then the level of real interest rates in the domestic country will be identical to the level of real interest rates in the foreign country. The proposition that real interest rates will converge to the same level across different markets is known as the **real interest rate parity** condition. This concept can be interpreted as an application of the law of one price to securities internationally:

$$r_f - r_d = 0$$

If real interest rates are equal across markets, such that $r_f - r_d = 0$, then it also follows that the foreign–domestic nominal yield spread will be solely determined by the foreign–domestic expected inflation differential:

$$i_f - i_d = \pi_f^e - \pi_d^e$$

This is known as the **international Fisher effect**.[14]

As noted earlier, the various parity relationships may seem of little empirical significance when examined over short time horizons. However, studies show that over longer time periods there is a discernible interaction among nominal interest rates, exchange rates, and inflation rates across countries, such that the international parity conditions described here serve as an anchor for longer-term exchange rate movements.

EXAMPLE 10-5 PPP, Real Exchange Rates, and the International Fisher Effect

An Australian-based fixed-income asset manager is deciding whether to allocate money between Australia and Japan. (As before, the AUD is the domestic currency.) Australia's one-year deposit rates are 5 percent, considerably higher than Japan's at 1 percent, but the Australian dollar is estimated to be roughly 10 percent overvalued relative to Japanese yen based on purchasing power parity. Before making her asset allocation, the asset manager considers the implications of interest rate differentials and PPP imbalances.

1. All else being equal, which of the following events would restore the Australian dollar to its PPP value?
   A. The Japanese inflation rate increases by 4 percent.
   B. The Australian inflation rate decreases by 10 percent.
   C. The JPY/AUD exchange rate declines by 10 percent.

[14]The reader should be aware that some authors refer to uncovered interest rate parity as the international Fisher effect. We reserve this term for the relationship between nominal interest rate differentials and expected inflation differentials because the original (domestic) Fisher effect is a relationship between interest rates and expected inflation.

2. If both the Australian and Japanese consumer price index (CPI) price levels increased by 5 percent, all else being equal, then the change in the real exchange rate $q_{JPY/AUD}$ would be *closest* to:
   A. 0 percent.
   B. 5 percent.
   C. 10 percent.
3. If real interest rates in Japan and Australia were equal, then under the international Fisher effect, the inflation rate differential between Japan and Australia would be *closest* to:
   A. 0 percent.
   B. 4 percent.
   C. 10 percent.
4. According to the theory and empirical evidence regarding purchasing power parity, which of the following would *not* be true if PPP holds in the long run?
   A. An exchange rate's equilibrium path should be determined by the long-term trend in domestic price levels relative to foreign price levels.
   B. Deviations from PPP might occur over short- and medium-term periods, but fundamental forces should eventually work to push exchange rates toward their long-term PPP path.
   C. High-inflation countries should tend to see their currencies appreciate over time.
5. Which of the following would *best* explain the failure of the absolute version of PPP to hold?
   A. Inflation rates vary across countries.
   B. Real interest rates are converging across countries.
   C. Trade barriers exist, and different product mixes are consumed across countries.

Solution to 1: C is correct. If the Australian dollar is overvalued by 10 percent on a PPP basis, with all else held equal, a depreciation of the JPY/AUD rate by 10 percent would move the Australian dollar back to equilibrium.

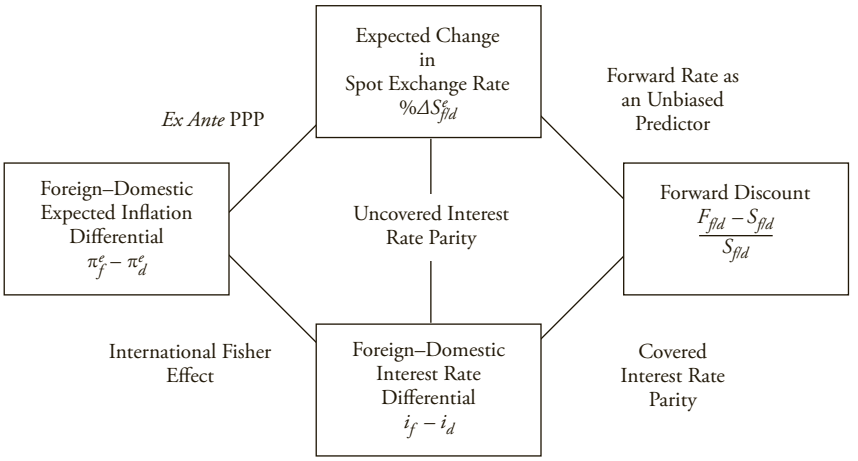
Solution to 2: A is correct. Based on our equation for the real exchange rate, these 5 percent changes would appear in both the numerator and denominator and cancel each other out.

Solution to 3: B is correct. If the real interest rates were equal, then the difference in nominal yields would be explained by the difference in inflation rates (5% – 1%).

Solution to 4: C is correct. According to PPP, high-inflation countries should see their currencies depreciate (at least over the longer term) in order to re-equilibrate real purchasing power between countries.

Solution to 5: C is correct. The absolute version of PPP assumes that all goods and services are identical, consumed in equal proportions across different countries, and freely tradable internationally.

EXHIBIT 10-4 How Spot Exchange Rates, Forward Exchange Rates, and Interest Rates Are Linked Internationally



Source: Rosenberg (2003).

3.1.6. International Parity Conditions: Tying All the Pieces Together

We now summarize the key international parity conditions and describe how they are all linked. We begin with the right-hand side of Exhibit 10-4. Starting in the lower right-hand corner, we note the following three points:

1. According to the theory of covered interest rate parity, arbitrage ensures that nominal interest rate spreads must equal the percentage forward premium (or discount).
2. According to the hypothesis of uncovered interest rate parity, the expected percentage appreciation (or depreciation) of the spot exchange rate should, on average, be reflected in the nominal interest rate spread.
3. If both covered and uncovered interest rate parity hold—that is, the nominal yield spread equals both the forward premium (or discount) and the expected percentage change in the spot exchange rate—then the forward exchange rate will be an unbiased predictor of the future spot exchange rate.

This brings us to the upper right-hand corner of Exhibit 10-4. Turning now to the left-hand side of the exhibit and beginning in the upper left-hand corner, we note three more points:

4. According to the *ex ante* PPP approach to exchange rate determination, the expected change in the spot exchange rate should equal the expected difference in domestic and foreign inflation rates.
5. Assuming the Fisher effect holds in each market—that is, the nominal interest rate in each market equals the real interest rate plus the expected inflation rate—and also assuming that real interest rates across all markets are broadly the same (real interest rate parity), then the nominal yield spread between domestic and foreign markets will equal the foreign–domestic expected inflation differential, which is the international Fisher effect.

6. If *ex ante* PPP and Fisher effects hold, then expected inflation differentials should equal both the expected change in the exchange rate and the nominal interest rate differential. This implies that the expected change in the exchange rate equals the nominal interest rate differential, which is uncovered interest rate parity.

Uncovered interest rate parity, at the center of the exhibit, brings us back to the right-hand loop (point 2) and, because covered interest rate parity is enforced by arbitrage (point 1), implies that the forward exchange rate is an unbiased predictor of the future spot exchange rate (point 3).

Exhibit 10-4 shows that if all the key international parity conditions held at all times, then the expected percentage change in the *spot* exchange rate would equal:

- The forward premium or discount (expressed in percentage terms).
- The nominal yield spread between countries.
- The difference in expected national inflation rates.

In other words, if all these parity conditions held, it would be impossible for a global investor to earn consistent profits on currency movements. If forward exchange rates accurately predicted the future path of spot exchange rates, there would be no way to make money in forward exchange speculation. If high-yield currencies fell in value versus low-yield currencies exactly in line with the path implied by nominal interest rate spreads, all markets would offer the same currency-adjusted total returns over time. Investors would have no incentive to shift funds from one market to another based solely on currency considerations.

EXAMPLE 10-6 The Relationships among the International Parity Conditions

1. Which of the following is a no-arbitrage condition?
   A. Real interest rate parity
   B. Covered interest rate parity
   C. Uncovered interest rate parity
2. Forward rates are unbiased predictors of future spot rates if two parity conditions hold. Which of the following is *not* one of these conditions?
   A. Real interest rate parity
   B. Covered interest rate parity
   C. Uncovered interest rate parity
3. The international Fisher effect requires all but which of the following to hold?
   A. *Ex ante* PPP
   B. Absolute PPP
   C. Real interest rate parity
4. The forward premium/discount is determined by nominal interest rate differentials because of:
   A. the Fisher effect.
   B. covered interest parity.
   C. real interest rate parity.

5. If all of the key international parity conditions held at all times, then the expected percentage change in the spot exchange rate would equal all *except* which of the following?
   A. The real yield spread
   B. The nominal yield spread
   C. The expected inflation spread

Solution to 1: B is correct. Covered interest rate parity is enforced by equating the investment return on two riskless investments (domestic and currency-hedged foreign investments).

Solution to 2: A is correct. Both covered and uncovered interest rate parity must hold for the forward rate to be an unbiased predictor of the future spot rate. Real interest rate parity is not required.

Solution to 3: B is correct. The international Fisher effect is based on real interest rate parity and *ex ante* PPP (not absolute PPP).

Solution to 4: B is correct. The forward/premium is determined by covered interest rate arbitrage.

Solution to 5: A is correct. If all the international parity conditions held, the real yield spread would equal zero regardless of expected changes in the spot exchange rate.

3.2. Assessing an Exchange Rate's Equilibrium Level

As noted earlier, there is evidence that PPP may be an appropriate framework with which to assess fair value in the FX markets over sufficiently long horizons. Nonetheless, the significant and persistent departures of observed exchange rates from such estimates of long-run equilibrium underscore the serious shortcomings of PPP as a real-time gauge of value. Hence, there have been many attempts to extend, supplement, or replace the parity-based framework for assessing a currency's short- or long-run equilibrium value.

Consider the dilemma facing FX forecasters and portfolio managers. Exchange rates tend to exhibit substantial noise over short time periods. From a strategic standpoint, an investor needs to be able to filter out such noise to get a better handle on the direction exchange rates are likely to take over the medium to long term. Knowing what constitutes a currency's real long-run equilibrium value would help investors better manage longer-term currency risk exposures.

From policy makers' standpoint, knowing what constitutes a currency's real long-run equilibrium value would enable them to better quantify the degree of exchange rate misalignment that may exist at any point in time. If there is an interest in maintaining a level playing field among countries that engage in international trade, then it is incumbent upon global officials to closely monitor whether some countries might be attempting to gain an unfair competitive advantage by keeping their currencies significantly undervalued relative to the currencies of their trading partners.

In recent decades, exchange rate policies have often been at the center of debates among G-20 policy makers. In the 1980s and 1990s, the value of the Japanese yen was a major point of contention in United States/Japan trade negotiations. In the 2000s, the value of the

Chinese yuan dominated United States/China trade talks. In response to the needs of its member countries, and in connection with its mandate to ensure the stability of the international monetary system, the International Monetary Fund (IMF) has been deeply involved in assessing the long-run value of exchange rates.

One of the IMF's core mandates is exchange rate surveillance. The IMF Consultative Group on Exchange Rate Issues (CGER) has used a three-pronged approach to derive long-run equilibrium exchange rate assessments for both developed and emerging market currencies:

1. The **macroeconomic balance approach** estimates how much exchange rates need to adjust in order to close the gap between the medium-term expectation for a country's current account imbalance and that country's normal (or sustainable) current account imbalance.
2. The **external sustainability approach** differs from the macroeconomic balance approach by focusing on stocks of outstanding assets or debt rather than on current account flows. The external sustainability approach calculates how much exchange rates would need to adjust to ensure that a country's net foreign-asset/GDP ratio or net foreign-liability/GDP ratio stabilizes at some benchmark level.
3. A reduced-form econometric model seeks to estimate the equilibrium path that a currency should take on the basis of the trends in several key macroeconomic variables, such as a country's net foreign asset position, its terms of trade, and its relative productivity.

The IMF finds that these three approaches are complementary in that they often generate similar assessments of fair value for most countries, although in some cases significant differences do arise.

The CGER estimates of long-run fair value help guide policy makers in determining how far exchange rates may be deviating from their long-run equilibrium levels. Interestingly, the IMF finds that the CGER estimates do have predictive value regarding future changes in *real* exchange rates. These methodologies frequently predict the *direction* of the change in the real exchange rates but unfortunately often miss in terms of *magnitude*. The IMF finds that its estimates of long-run fair value tend to overpredict both the future appreciation of undervalued currencies and the future depreciation of overvalued currencies. That is, currencies that are estimated to be significantly misaligned by the IMF tend to remain significantly misaligned in the future, although the magnitude of the misalignment tends to shrink over time.

These three IMF approaches can be viewed as an attempt to expand beyond the international parity conditions in order to explain some of the *mechanisms* that help bring exchange rates back toward their long-term equilibrium values. Exchange rates reflect the sum of all forces influencing trade and capital flows. Loosely speaking, the current account reflects the flows in the real economy while the capital account reflects financial flows. Because the balance of payments accounts must always balance, the current account must be matched by an equal and opposite balance in the capital account. This balance is achieved primarily by movements in exchange rates, which bring all the actions of the wide diversity of currency market participants into alignment. The IMF's three approaches largely represent different perspectives on this single equilibrating mechanism. Put somewhat simplistically, the macroeconomic balance approach focuses on the flows needed to achieve long-term equilibrium in the current account. (A country running a persistent current account deficit will eventually need to see its currency depreciate in order to restore its trade competitiveness.) In turn, the external sustainability approach focuses on the stock of net external debt that leads to long-term equilibrium in the capital account. As a hybrid approach, the reduced-form

econometric model combines elements of both the current and capital accounts in a single, statistical equation.

In the following section, we derive a model of exchange rates that encapsulates both the long-run equilibrium value of a currency as well as some of the forces that drive exchange rate movements around this equilibrium level.

EXAMPLE 10-7 Long-Run Equilibrium Exchange Rates

Various methods exist to calculate long-run equilibrium values for exchange rates, and the competing approaches do not always agree on what exchange rate level constitutes long-run equilibrium. The following matrix focuses on two popular approaches to assess long-run fair value—PPP and the macroeconomic balance approach. Currencies that are out of line with fair value are likely to be characterized by one of the four cells (I, II, III, or IV) in the matrix.

Assessing Long-Run Fair Value in the FX Markets Combinations of PPP Misalignment and Current Account Imbalance

	PPP Overvaluation	PPP Undervaluation
Unsustainably large current account surplus	I	III
Unsustainably large current account deficit	II	IV

1. If a currency is described by cell II, which of the following would *best* help bring the exchange rate closer to long-run fair value?
   A. Pursue policies to encourage currency depreciation.
   B. Pursue policies to encourage currency appreciation.
   C. The recommended course of action is unclear due to different recommended courses of action coming from PPP and current account balance considerations.
2. If a currency is described by cell III, which of the following would *best* help bring the exchange rate closer to long-run fair value?
   A. Pursue policies to encourage currency depreciation.
   B. Pursue policies to encourage currency appreciation.
   C. The recommended course of action is unclear due to different recommended courses of action coming from PPP and current account balance considerations.
3. If a currency is described by cell IV, which of the following would *best* help bring the exchange rate closer to long-run fair value?
   A. Pursue policies to encourage currency depreciation.
   B. Pursue policies to encourage currency appreciation.
   C. The recommended course of action is unclear due to different recommended courses of action coming from PPP and current account balance considerations.

Solution to 1: A is correct. The currency is overvalued on a PPP basis (it should depreciate), and to obtain a sustainable current account, the domestic currency must depreciate against foreign currencies.

Solution to 2: B is correct. The currency is undervalued on a PPP basis (it should appreciate), and to obtain a sustainable current account, the domestic currency must appreciate against foreign currencies.

Solution to 3: C is correct. The currency is undervalued on a PPP basis, but to correct the current account deficit, it should depreciate.

3.3. Tying It Together: A Model That Includes Long-Term Equilibrium

Given the range of competing methodologies and the judgments that must be made in implementing them, it should not come as a surprise that there is often significant disagreement among both economists and market participants as to the appropriate fair value level for an exchange rate.

However, the attempt to define the long-run equilibrium value of a currency is not without merit. Most economists believe that there is an equilibrium level or a path to that equilibrium value that a currency will gravitate toward in the long run. Although medium-term cyclical deviations from the long-run equilibrium path can be both sizable and persistent, over the long run, fundamental forces should eventually drive the currency back toward its long-run equilibrium path. Evidence for both developed and emerging market economies suggests that misalignments tend to build up gradually over time. As these misalignments build, they are likely to generate serious economic imbalances that will eventually lead to correction of the underlying exchange rate misalignment.

At this stage, it is useful to develop a simple mathematical model of long-run exchange rate determination that will be helpful in subsequent discussions. The model incorporates the notion of long-run convergence of the real exchange rate to fair value and also demonstrates the role of real interest rate differentials. As before, to simplify the explanation and link it to our previous work, let's continue to use the foreign/domestic currency notation (*f/d*).

The model combines two assumptions. First, uncovered interest rate parity holds in the long run, implying that over a sufficiently long horizon,

$$\%\Delta S^{e}_{f/d} = i_f - i_d \tag{10-6}$$

From the definition of the real exchange rate, we also know that:

$$\%\Delta S^{e}_{f/d} = \%\Delta q^{e}_{f/d} - (\pi^{e}_{d} - \pi^{e}_{f}) \tag{10-7}$$

By subtracting Equation 10-6 from Equation 10-7, we can show that the expected change in the real exchange rate ($\%\Delta q^{e}$) must equal the real interest rate differential ($r_f - r_d$):

$$\%\Delta q^{e}_{f/d} = (i_f - \pi^{e}_{f}) - (i_d - \pi^{e}_{d}) \tag{10-8}$$

$$\%\Delta q^{e}_{f/d} = (r_f - r_d) \tag{10-9}$$

The second assumption is that the real exchange rate is expected to converge to its long-run equilibrium value, $\overline{q}_{f/d}$. The long-run equilibrium may reflect PPP, one of the methods employed by the IMF, or some combination of estimates. Formally, we assume:[15]

$$\%\Delta q^{\varepsilon}_{f/d} = \overline{q}_{f/d} - q_{f/d} \tag{10-10}$$

Combining Equations 10-9 and 10-10 gives:

$$q_{f/d} = \overline{q}_{f/d} + (r_d - r_f) \tag{10-11}$$

The model can be extended to incorporate a risk premium (φ) that investors require in order to hold each country's securities. Among other factors, the required risk premium likely reflects the perceived sustainability of the country's external balances. Incorporating relative risk premiums into the model yields:

$$q_{f/d} = \overline{q}_{f/d} + [(r_d - r_f) - (\varphi_d - \varphi_f)] \tag{10-12}$$

According to Equation 10-12, the level of the real exchange rate can be expressed as a function of three key variables: (1) the level of the real long-run equilibrium exchange rate, (2) the real interest rate differential, and (3) relative risk premiums. With all else held equal, a currency's real value would be expected to rise over time in response to (1) an upward revision in that currency's real long-term equilibrium value, (2) a rise in domestic relative to foreign real interest rates, and/or (3) a decline in domestic relative to foreign risk premiums.

4. THE CARRY TRADE

According to uncovered interest rate parity, high-yield currencies are expected to depreciate in value, while low-yield currencies are expected to appreciate in value. If uncovered interest rate parity held at all times, investors would not be able to profit from a strategy that undertook long positions in baskets of high-yield currencies and short positions in baskets of low-yield currencies. The change in spot rates over the tenor of the forward contracts would cancel out the interest rate differentials locked in at the inception of the position.

Uncovered interest rate parity is one of the most widely tested propositions in the field of international finance. Literally hundreds, if not thousands, of academic studies have tested whether uncovered interest rate parity has held for both developed and emerging market (EM) countries. Overwhelmingly, the evidence suggests that uncovered interest rate parity does not hold, at least over short- and medium-run time periods. These studies have generally found

[15]The very careful reader will note that the two sides of Equation 10-10 involve different units—the left-hand side is a percentage, while the right-hand side is in foreign currency units. Dividing the right hand side of Equation 10-10 by $q_{f/d}$ would make the units consistent. With this change, Equation 10-11 would become $q_{f/d} = \overline{q}_{f/d} + q_{f/d}(r_d - r_f)$. For simplicity, we will work with the slightly simpler versions of these equations given in the text.

that high-yield currencies, on average, have not depreciated, and low-yield currencies have not appreciated, to the levels predicted by interest rate differentials.

These findings underscore the potential profitability of a trading strategy known as the **FX carry trade**, which involves taking on long positions in high-yield currencies and short positions in low-yield currencies. The latter are often referred to as "funding currencies." Historical evidence shows that such carry trade strategies often generate attractive excess returns over extended periods. One argument for the persistence of the carry trade is that the yields in higher interest rate countries are a risk premium for a more unstable economy, while low-yield currencies represent less risky markets. Although history has demonstrated that carry trades generally earn positive excess returns in most market conditions, elevated levels of volatility and/or perceived risk in the financial markets can turn these excess returns into substantial losses very quickly. That is, small increases in asset market and/or FX volatility are unlikely to materially affect the positive excess returns earned on carry strategies. But during turbulent periods when asset price volatility and/or FX volatility rise sharply, the realized returns on long high-yield currency positions will tend to decline dramatically, while the realized returns on low-yield currencies will tend to rise just as dramatically.

To understand why, we need to recognize the nature of the risk and reward on the carry trade. The reward is the gradual accrual of the interest rate differential—essentially a flow of income that is unrelated to exchange rate volatility. The risk arises from the potential for adverse exchange rate movements—capital losses can occur virtually instantaneously. During periods of low turbulence, investors may feel relatively confident that exchange rate movements will not jeopardize the gradual accrual of the interest rate differential. In periods of high volatility, however, the risk of an adverse exchange rate movement rises sharply relative to the gradual flow of income, and investors may rush to unwind the carry trade. In the process, they sell the high-yielding currency basket and buy the low-yielding one, shifting realized returns sharply in favor of the low-yielding basket. The upshot is that during periods of low volatility, carry trades tend to generate positive excess returns, but they are prone to significant crash risk in turbulent times.

The tendency for carry trades to experience periodic crashes reveals itself in the distribution of carry trade returns. Academic studies find that the distribution of returns for G-10 and EM carry trades do not conform to a normal distribution. Rather, the distributions tend to be more peaked and to have fatter tails that are negatively skewed. The more peaked distribution around the mean implies that carry trades have typically generated a larger number of trades with small gains or losses than would occur with the normal distribution. The negative skew and fat tails indicate that carry trades have tended to have more frequent and larger losses than would have been experienced had the distribution of returns been normal. That is, even though carry trades have generated positive excess returns on average in the past, the presence of a large negative skew means that such strategies have been exposed to crash risk.

The primary reason for these crash risks relates to the fact that the carry trade is by nature a leveraged position; investors borrow in the funding currency and invest the funds in a high-yielding currency. Like all leverage, this magnifies the impact of losses and gains relative to the investors' equity base. Moreover, because low-volatility regimes have tended to be the norm and often last for extended periods, investors can become complacent, allowing carry trade positions to grow to increasingly large levels in a search for yield. This crowded positioning tends to unwind rapidly when a shock to the market occurs, as many traders try

to exit their position almost simultaneously before the leverage effects wipe out their equity. Another factor that accelerates the selling is that traders often leave stop-loss orders with their brokers that are triggered when price declines reach a certain level. These combined effects can lead to cascades of selling in which position liquidation begets further position liquidation. Finally, during periods of market turmoil, there is generally a flight to safety into those assets and currencies that seem to offer the most protection during times of uncertainty. This has typically favored low-interest-rate currencies (i.e., popular funding currencies) such as the Japanese yen and the Swiss franc, which are typically viewed as havens in times of market stress.

Recent research has focused on how one could best manage the downside crash risks associated with carry trades. One particular approach advocated by a number of investment banks and academics is to use a volatility filter to determine whether carry trade positions should be left open or closed. Under this approach, if the average level of FX volatility (or some other measure of market turbulence such as equity market volatility) were to trade below a specified threshold in the options market, then a signal would be generated to open carry trade positions. Likewise, if FX volatility or equity market volatility were to rise above a higher threshold, then a signal would be generated that those positions should be closed or reversed.

Another approach recommended by some observers is to combine valuation and carry into a single integrated strategy. Under this approach, purchasing power parity benchmarks can be used to determine whether carry trade positions should be opened or closed. That is, a valuation overlay would recommend that high yielders be overweighted and low yielders be underweighted when exchange rates lie inside prescribed PPP bands. If, instead, one or more of the high yielders were to become overvalued relative to their prescribed PPP threshold bands, the likelihood of a correction would increase and a valuation overlay would recommend closing or reversing the high-yield position. The opposite case would apply for low yielders.

These and other risk management tools are not foolproof, however, and investors need to be mindful of the considerable negative tail risks associated with carry trades.

From the perspective of policy makers, there is a clear concern that carry trade activities might be playing a major role in generating exchange rate misalignments around the world. As such activities have become a more important part of the FX landscape, there is a risk that a global search for yield could drive high-yield currencies deep into overvalued territory, which could have serious negative consequences for economic activity in high-yield markets. In such an environment, one runs the risk that the monetary authorities in high-yield markets might feel compelled to stem the inflow of capital into their market—for example, by introducing capital controls—to help prevent an unwarranted or undesired appreciation of their currencies.

Another policy-related danger of carry trade activities is that if a global search for yield encourages international investors to take on highly leveraged positions in carry trades, and if speculative positions begin to lean too heavily in one direction, a forced unwinding of highly leveraged carry trade positions could precipitate a serious currency or financial crisis. The carry trade unwind of 2008 illustrates the impact that these trades could have on exchange rates. During that period, high-yield developed market currencies such as the Australian dollar and the New Zealand dollar—as well as many high-yielding emerging market currencies—lost considerable ground, even though none of those high-yielding markets were at the epicenter of the 2007–2009 global financial crisis.

EXAMPLE 10-8 Carry Trade Strategies

A currency fund manager is considering allocating a portion of her FX portfolio to carry trade strategies. The fund's investment committee asks the manager a number of questions about why she has chosen to become involved in FX carry trades and how she will manage the risk of potentially large downside moves associated with the unwinding of carry trades. Which of the following would be her *best* responses to the investment committee's questions?

1. Carry trades can be profitable when:
   A. covered interest rate parity does not hold.
   B. uncovered interest rate parity does not hold.
   C. the international Fisher effect does not hold.
2. Over time, the return distribution of the fund's FX carry trades is *most likely* to resemble:
   A. a normal distribution with fat tails.
   B. a distribution with fat tails and a negative skew.
   C. a distribution with thin tails and a positive skew.
3. The volatility of the fund's returns relative to its equity base is *best* explained by:
   A. leverage.
   B. low deposit rates in the funding currency.
   C. the yield spread between the high- and low-yielding currencies.
4. The risk management strategy that is *most* likely to reduce some of the negative tail risk associated with FX carry trades is to:
   A. use forward contracts to sell the funding currency.
   B. exit the carry trade position when implied FX volatility drops below a certain threshold.
   C. exit the carry trade when the funding currency drops below its PPP level.
5. A Tokyo-based asset manager enters into a carry trade position based on borrowing in yen and investing in one-year Australian LIBOR.

Today's One-Year LIBOR		Currency Pair	Spot Rate Today	Spot Rate One Year Later
JPY	0.10%	JPY/USD	81.30	80.00
AUD	4.50%	USD/AUD	1.0750	1.0803

After one year, the all-in return to this trade, measured in JPY terms, would be *closest* to:

A. +1.84 percent.
B. +3.23 percent.
C. +5.02 percent.

Solution to 1: B is correct. The carry trade is based on the supposition that uncovered interest rate parity does not hold.

Solution to 2: B is correct. The crash risk of carry trades implies a fat-tailed distribution skewed toward a higher probability of large losses (compared with a normal distribution).

Solution to 3: A is correct. Carry trades are leveraged trades (borrow in the funding currency, invest in the high-yield currency), and leverage increases the volatility in the investor's return on equity.

Solution to 4: C is correct. One would want to exit the carry trade position when the likelihood of either depreciation of the high-yield currency or appreciation of the low-yield currency is building. When the funding currency is undervalued on a PPP basis, the likelihood that it might appreciate is increasing.

Solution to 5: B is correct. To calculate the all-in return to a Japanese investor in a one-year AUD LIBOR deposit, we must first calculate the current and one year later JPY/AUD cross-rates. Because one USD buys JPY81.30 today and one AUD buys USD1.0750 today, today's JPY/AUD cross-rate is the product of these two numbers: $81.30 \times 1.0750 = 87.40$ (rounding to two decimal places). Similarly, one year later the observed cross-rate is $80.00 \times 1.0803 = 86.42$ (rounded to two decimal places). Accordingly, measured in yen, the investment return for the unhedged Australian LIBOR deposit is *closest* to:

$$\frac{1}{87.40}(1 + 4.50\%)86.42 = (1.0333)$$

Against this 3.33 percent *gross* return, however, the manager must charge the borrowing costs to fund the carry trade investment (one-year yen LIBOR was 0.10 percent). Hence, the *net* return on the carry trade is closest to 3.33% – 0.10% = 3.23%.

5. THE IMPACT OF BALANCE OF PAYMENTS FLOWS

The current account balance of a country represents the sum of all recorded transactions in traded goods, services, income, and net transfer payments in a country's overall balance of payments. Countries that run persistent current account deficits often see their currencies depreciate over time. Similarly, countries that run persistent current account surpluses often see their currencies appreciate over time.

Because the balance of payments accounts must always balance, the current account must be matched by an equal and opposite balance in the capital account.[16] Loosely speaking, the current account reflects the flows in the real economy, while the capital account reflects financial flows. Although this equality must always hold, it is not inconsequential. Decisions with respect to trade flows (the current account) and investment/financing flows (the capital account) are typically made by different entities with different perspectives and motivations.

[16]The official balance of payments accounts make a distinction between the capital account and the financial account based on the nature of the assets involved. For simplicity, we use the term *capital account* here to reflect all investment/financing flows.

Their decisions are brought into alignment by changes in market prices and/or quantities. One of the key prices, perhaps *the* key price, in this process is the exchange rate.

It turns out that *investment/financing decisions are usually the dominant factor in determining exchange rate movements, at least in the short to intermediate term*. There are four main reasons for this. First, prices of real goods and services tend to adjust much more slowly than exchange rates and other asset prices. Second, production of real goods and services takes time, and demand decisions are subject to substantial inertia. In contrast, liquid financial markets allow virtually instantaneous redirection of financial flows. Third, whereas current spending/production decisions reflect only purchases/sales of current production, investment/financing decisions reflect not only the financing of current expenditures but also reallocation of existing portfolios. Fourth, *expected* exchange rate movements can induce very large short-term capital flows. This tends to make the *actual* exchange rate very sensitive to the currency views held by owners and managers of liquid assets.

In this section, we first examine the impact of current account imbalances on exchange rates. Then we take a closer look at capital flows.

5.1. Current Account Imbalances and the Determination of Exchange Rates

Current account trends influence the path of exchange rates over time through several mechanisms:

- The flow supply/demand channel.
- The portfolio balance channel.
- The debt sustainability channel.

We briefly discuss each of these in the following pages.

5.1.1. The Flow Supply/Demand Channel

The flow supply/demand channel is based on a fairly simple model that focuses on the fact that purchases and sales of internationally traded goods and services require the exchange of domestic and foreign currencies in order to arrange payment for those goods and services. For example, if a country sold more goods and services than it purchased (i.e., the country was running a current account surplus), then the demand for its currency should rise, and vice versa. Such shifts in currency demand would tend to exert upward pressure on the value of the surplus nation's currency and downward pressure on the value of the deficit nation's currency.

Hence, we would expect to find that countries with persistent current account surpluses would see their currencies appreciate over time, and vice versa. The question remains whether such trends can go on indefinitely. At some point, domestic currency strength should contribute to deterioration in the trade competitiveness of the surplus nation while domestic currency weakness should contribute to an improvement in the trade competitiveness of the deficit nation. Thus, the exchange rate responses to these surpluses and deficits should eventually help eliminate—in the medium to long run—the source of the initial imbalances.[17]

[17]Currency depreciation will improve (and appreciation will worsen) the current account if the generalized Marshall–Lerner condition is satisfied. This condition was developed and discussed in the preceding chapters.

The amount by which exchange rates must adjust to restore current accounts to balanced positions depends on a number of factors:

- The initial gap between imports and exports.
- The response of import and export prices to changes in the exchange rate.
- The response of import and export demand to the change in import and export prices.

Regarding the first factor, when the initial gap between imports and exports is relatively wide for a deficit nation, export growth would need to far outstrip import growth in percentage terms in order to narrow the current account deficit. Unless export demand is far more price elastic than import demand, a large initial deficit may require a substantial depreciation of the currency to bring about a meaningful correction of the trade imbalance.

Normally, a depreciation of the deficit country's currency should result in an increase in import prices in domestic currency terms and a decrease in the deficit country's export prices in foreign currency terms. However, empirical studies often find limited pass-through effects of exchange rate changes into traded goods prices. For example, many studies find that for every 1 percent decline in a currency's value, import prices rise by only 0.5 percent, and in some cases by even less because foreign producers tend to lower their profit margins in an effort to preserve market share. In light of the limited pass-through of exchange rate changes into traded goods prices, the exchange rate adjustment required to narrow a trade imbalance may be far larger than would otherwise be the case.

Even if traded goods prices ultimately adjust one for one with the change in exchange rates, the response of import and export demand to those price changes might not be sufficient to correct a sizable current account imbalance. In fact, many empirical studies find that the response of import and export demand to changes in traded goods prices is often quite sluggish. As a result, relatively long lags, lasting several years, can occur between (1) the onset of exchange rate changes and (2) the ultimate adjustment in traded goods prices, and then (3) the eventual impact of those price changes on import demand, export demand, and the underlying current account imbalance.

5.1.2. The Portfolio Balance Channel

The second mechanism through which current account trends influence exchange rates is the so-called portfolio balance channel. Current account imbalances shift financial wealth from deficit nations to surplus nations. Over time, this may lead to shifts in global asset preferences, which in turn could exert a marked impact on the path of exchange rates. For example, nations running large current account surpluses versus the United States might find that their holdings of U.S. dollar-denominated assets exceed the amount they desire to hold in a portfolio context. Attempts to reduce their dollar holdings to desired levels could then have a profound negative impact on the dollar's value.

5.1.3. The Debt Sustainability Channel

The third mechanism through which current account surpluses can affect exchange rates is the so-called debt sustainability channel. According to this channel, there should be some upper limit on the ability of countries to run persistently wide current account deficits. If a country runs a large and persistent current account deficit over time, eventually it will experience an unending rise in debt owed to foreign investors. If such investors believe that the deficit country's external debt is rising to unsustainable levels, they are likely to reason that a major

depreciation of the deficit country's currency will be required at some point to ensure that the current account deficit narrows significantly and that the external debt stabilizes at a level deemed sustainable.

The existence of persistent current account imbalances will tend to alter the market's notion of what exchange rate level represents the true, real long-run equilibrium value. For deficit nations, ever-rising net external debt levels as a percentage of GDP should give rise to steady (but not necessarily smooth) downward revisions in market expectations of the currency's real long-run equilibrium value. For surplus countries, ever-rising net external asset levels as a percentage of GDP should give rise to steady upward revisions of the currency's real long-run equilibrium value. Hence, one would expect currency values to move broadly in line with trends in debt or asset accumulation.

Note that the debt sustainability channel is essentially the perspective underlying the IMF's external sustainability approach to determining the long-run equilibrium exchange rate.

Persistent Current Account Deficits: The U.S. Current Account and the USD

The historical record indicates that the trend in the U.S. current account has been an important determinant of the long-term swings in the U.S. dollar's value but also that there can be rather long lags between the onset of a deterioration in the current account balance and an eventual decline in the dollar's value. For example, the U.S. current account balance deteriorated sharply in the first half of the 1980s, yet the dollar soared over that period. The reason for the dollar's strength over that period was that high U.S. real interest rates attracted large inflows of capital from abroad, which pushed the dollar higher despite the large U.S. external imbalance. Eventually, however, concerns on the part of market participants and policy makers regarding the sustainability of the ever-widening U.S. current account deficit triggered a major dollar decline in the second half of the 1980s.

History repeated itself in the second half of the 1990s, with the U.S. current account deficit once again deteriorating, yet the dollar soared over the same period. This time, the dollar's strength was driven by strong foreign direct investment and equity-related flows into the United States. Beginning in 2001, however, the ever-widening U.S. current account deficit, coupled with a decline in U.S. interest rates, made it more difficult for the United States to attract the foreign private capital needed to finance its current account deficit. The dollar eventually succumbed to the weight of ever-larger trade and current account deficits and began a multiyear slide starting in 2002–2003. Interestingly, the U.S. dollar has undergone three major downward cycles since the beginning of floating exchange rates: 1977–1978, 1985–1987, and 2002–2008. In each of those downward cycles, the dollar's slide was driven in large part by concerns over outsized U.S. current account deficits that were coupled with relatively low nominal and/or real short-term interest rates in the United States, which made it difficult to attract sufficient foreign capital to the United States to finance those deficits.

Exchange Rate Adjustment in Surplus Nations—Japan and China

Japan and, more recently, China represent examples of countries with large current account surpluses and the pressure that those surpluses can bring to bear on their currencies. In the case of Japan, the rising trend in its current account surplus has tended to exert persistent upward pressure on the yen's value versus the dollar over time. Part of this upward pressure has simply reflected the increase in demand for yen to make payment for Japan's merchandise exports. But some of the upward pressure on the yen might also have stemmed from rising commercial tensions between the United States and Japan.

Protectionist sentiment in the United States rose steadily in line with the rising bilateral trade deficit that the United States ran with Japan in the postwar period. U.S. policy makers contended that the yen was undervalued and needed to rise. With the rising Japan–U.S. bilateral trade imbalance contributing to more heated protectionist rhetoric, Japan felt compelled to tolerate steady, upward pressure on the yen to placate U.S. demands for further exchange rate adjustments. As a result, the yen's value versus the dollar has tended to move in sync with the trend in Japan's current account surplus.

In recent years, U.S. protectionist rhetoric has shifted to China. Given China's growing current account surplus, the Chinese currency has become a major point of contention between U.S. and Chinese authorities. The exchange rate regime between the Chinese yuan and the U.S. dollar can be characterized as a crawling peg in which the rate is allowed to fluctuate within a narrow band (currently ± 1 percent) around a fixing level determined by the Chinese government. This fixing rate varies from day to day. In recent years, China has engineered a gradual appreciation of its currency. Nonetheless, the Chinese authorities have intervened aggressively to moderate the rate of appreciation, accumulating massive foreign exchange reserves in the process, and many economists argue that significant upward adjustment in the Chinese yuan's value is still needed to narrow China's large current account surplus.

5.2. Capital Flows and the Determination of Exchange Rates

Greater financial integration of the world's capital markets and the increased freedom of capital to flow across national borders have increased the importance of global financial flows in determining exchange rates, interest rates, and broad asset price trends. One can cite a litany of examples in which global financial flows either caused or contributed to overshooting exchange rates, interest rates, or asset price bubbles. Among them are the following:

- As described in the box "Exchange Rate Adjustment in Surplus Nations—Japan and China," the dramatic rise in the U.S. dollar in the first half of the 1980s and again in the second half of the 1990s was powered by a significant rise in global demand for U.S. financial assets. The 1980s episode was driven by a major widening in yield spreads favoring

the United States, while the 1990s rise was powered in part by a significant rise in global demand for U.S. equities during the new economy/tech boom.

- Yen assets underperformed U.S. dollar assets over much of the 1995–2007 period, a 12-year span of ultralow Japanese short-term interest rates that gave rise to what became known as the "yen carry trade" as both Japanese and global fund managers borrowed in yen and invested the proceeds in higher-yielding assets in other markets. Such actions helped push the value of the yen significantly lower on a trend basis. Periodically, however, such positions became overextended and vulnerable to sudden reversals. In the fall of 1998, a major unwinding of the yen carry trade led to the collapse of several major hedge funds.
- Australian and New Zealand dollar-denominated assets have both significantly outperformed U.S. dollar-denominated assets over the past two decades, driven in large measure by capital inflows attracted by the persistently higher short-term interest rates in those markets relative to the rest of the industrial world.
- In numerous cases, global capital flows have helped fuel boomlike conditions in emerging market economies for a while, and then suddenly, often without adequate warning, those flows stopped and then reversed—with the reversal often causing a major economic downturn, a possible sovereign default, a serious banking crisis, or a significant currency depreciation. A recent IMF study of 109 episodes of major capital inflow surges over the 1987–2007 period found that, in more than one-third of the cases, the inflows stopped abruptly.

Excessive surges in capital inflows to emerging markets have often planted the seeds of an economic or currency crisis by contributing to (1) an unwarranted real appreciation of the emerging market currency; (2) a huge buildup in external indebtedness by emerging market governments, businesses, or banks; (3) a financial asset or property market bubble; (4) a consumption binge that contributed to explosive growth in domestic credit and/or the current account deficit; or (5) an overinvestment in risky projects and questionable activities.

Because episodes of capital flow surges to emerging markets have often ended badly, emerging market policy makers have made a concerted effort in recent years either to resist such inflows through the use of capital controls or to prevent capital inflows from pushing currency values to overvalued levels by intervening more heavily in the foreign exchange market. In the following analysis, we discuss the empirical evidence on the success, or lack thereof, of policy makers counteracting the negative consequences of capital flow surges.

5.2.1. Real Interest Rate Differentials, Capital Flows, and the Exchange Rate

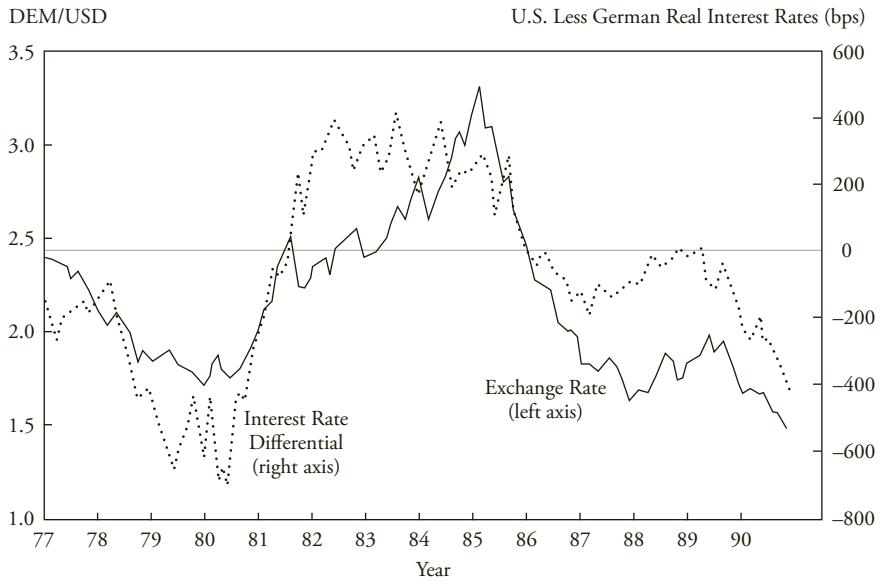
In Section 3.3, we developed a model in which real exchange rate movements around the long-run equilibrium are driven primarily by the response of capital flows to interest rate differentials, risk premiums, and expectations with respect to exchange rate movements themselves. For ease of reference, we repeat Equation 10-12 here:

$$q_{f/d} = \overline{q}_{f/d} + [(r_d - r_f) - (\varphi_d - \varphi_f)]$$

Recall that $q_{f/d}$ is the real exchange rate, r_d and r_f are real interest rates, φ_d and φ_f are risk premiums, and a bar over a variable indicates its long-run equilibrium value. Based on the model, movements in real interest rate and risk premiums differentials between countries can cause movements in the real exchange rate $q_{f/d}$ around its long-run equilibrium value $\overline{q}_{f/d}$.

This simple model can explain a wide range of phenomena such as (1) long-run cyclical trends in the U.S. dollar value, (2) the persistent excess returns that carry trade strategies have tended to offer, and (3) the impact of temporary bouts of risk aversion on exchange rates.

EXHIBIT 10-5 Deutsche Mark/USD Exchange Rate and U.S./German Real Interest Rate Differentials (10-Year Bond Yields less CPI, 1977–1990)

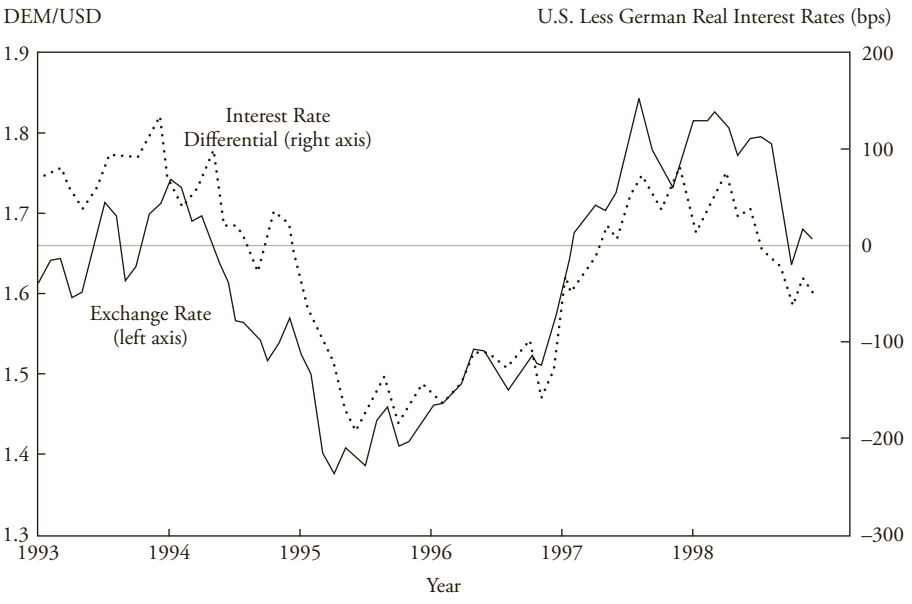


Source: Datastream.

The trend in real interest rate differentials played a pivotal role in driving the U.S. dollar's value during several major exchange rate cycles. As shown in Exhibit 10-5, the decline of the dollar in the late 1970s, the dramatic rise in its value in the first half of the 1980s, and its subsequent decline in the second half of the 1980s can be explained, to a large extent, by changes in U.S.–foreign real yield spreads. (In Exhibits 10-5 and 10-6, DEM/USD indicates the number of German deutsche marks per U.S. dollar.) The dollar's surge in the 1990s can also be explained, in part, by trends in real yield spreads. Exhibit 10-6 shows that the dollar's decline in the first half of the 1990s coincided with a significant narrowing in U.S.–foreign real yield spreads, while the dollar's subsequent rise in the second half of the 1990s coincided with a significant widening in U.S.–foreign real yield spreads. To be sure, other forms of capital movements were at work during these dollar cycles—notably foreign direct investment and equity-related portfolio flows, particularly in the second half of the 1990s—but Exhibits 10-5 and 10-6 suggest that relative interest rate trends played a major role in each of the dollar's major upward and downward moves.

Interest rate differentials can play a major role in driving the relative total return performance of competing markets, even if interest rate spreads have little impact on the trend in exchange rates. Consider the case of the Turkish lira versus the U.S. dollar. The lira attracted a lot of interest on the part of global fund managers over the 2002–2010 period, in large part because of its attractive yield levels. Turkish–U.S. short-term yield spreads averaged over 1,000 basis points during much of this period. As capital flowed into Turkey, the Turkish authorities intervened in the foreign exchange market in an attempt to keep the value of the lira broadly stable. As a result, international investors were not able to reap much in terms of currency gains over this period. Nevertheless, international investors were able to capture the high yields that Turkey offered, and over time, the cumulative effect was a significant outperformance of Turkish over U.S. assets

EXHIBIT 10-6 Deutsche Mark/USD Exchange Rate and U.S./German Real Interest Rate Differentials (10-Year Bond Yields less CPI, 1993–1998)



Source: Datastream.

during much of the 2002–2010 period. Indeed, a long Turkish lira/short U.S. dollar carry trade position generated significant long-run excess returns, mostly from the accumulated yield spread, while the return from the movement in the spot exchange rate was fairly small.

5.2.2. Interest Rate Differentials, Carry Trades, and Exchange Rates

Carry trades have generated significant positive excess returns for substantial periods in cases involving G-10 currencies as well as emerging market currencies. Sometimes those persistently positive excess returns show up in movements in the spot exchange rate, and sometimes the persistent returns show up in the contribution that cumulative yield spreads offer. Quite often, a combination of both factors jointly works to drive persistent positive excess returns. Equation 10-12 provided a useful framework to understand how each of those factors has come into play in driving international capital flows toward markets offering higher yields.

To understand why a high-yield country might attract significant inflows of capital that could exert considerable upward pressure on the high-yield currency's value for persistent periods of time, it is useful to rewrite the equation letting H and L denote the high-yield and low-yield currencies, respectively, with the high-yield currency as the base currency in the exchange rate:

$$q_{L/H} = \overline{q}_{L/H} + (r_H - r_L) - (\varphi_H - \varphi_L) \tag{10-13}$$

As a further step, we can break the real interest rate spread into its two constituent parts—the nominal interest rate spread and the expected inflation differential between the high-yield and low-yield countries:

$$q_{L/H} = \overline{q}_{L/H} + (i_H - i_L) - (\pi^{\varepsilon}_H - \pi^{\varepsilon}_L) - (\varphi_H - \varphi_L) \tag{10-14}$$

According to Equation 10-14, a high-yield currency's real value will tend to rise in the long run when the long-run equilibrium value of the high-yield currency ($\overline{q}_{L/H}$) is trending higher. Typically, however, oscillations around this long-run equilibrium trend can persist for relatively long periods of time (refer back to Exhibit 10-2 for an illustrative diagram of this process). These cyclical movements around the long-run equilibrium trend are often associated with movements in the differentials in Equation 10-14. That is, the real exchange rate $q_{L/H}$ will tend to rise, relative to its long-run equilibrium value $\overline{q}_{L/H}$, when (1) the nominal yield spread between the high- and low-yield market rises, (2) inflation expectations in the high-yield market decline relative to the low-yield market, and/or (3) the risk premium demanded by investors to hold the assets of the high-yield market declines relative to the low-yield market. It is important to recognize that movements in all of these various differentials can be gradual but persistent and thereby lead to persistent movements in the exchange rate.

Consider the case of a high-yield, inflation-prone emerging market country that wants to promote price stability and long-term sustainable growth. To achieve price stability, policy makers in the high-yield economy will initiate a tightening in monetary policy by gradually raising the level of domestic interest rates relative to yield levels in the rest of the world. Assuming that the tightening in domestic monetary policy is sustained, inflation expectations in the high-yield economy should gradually decline relative to the trend in inflation expectations in the rest of the world. The combination of sustained wide nominal yield spreads and a steady narrowing in relative inflation expectations should exert upward pressure on the real yield spread and thus on the high yield currency's value.

To achieve their long-run objectives of price stability and sustainable economic growth, policy makers in high-yield markets may not rely solely on a tight monetary policy. To reinforce initiatives on the monetary policy front, government authorities might also pursue policies that promote lower government deficits, liberalize financial markets, remove capital flow restrictions, attract foreign direct investment, promote privatization, and/or encourage better business practices. By encouraging an economic environment that promotes price stability, long-run growth, and a stronger and more stable financial system, such policies should also encourage investors to gradually demand a lower risk premium to hold the high-yield currency's assets.

All of these various government measures should gradually work to boost the long-run competitiveness of the high-yield country and, in the process, should lead market participants to gradually revise upward their assessment of the long-run equilibrium value of that country's currency ($\overline{q}_{L/H}$). And, as Equation 10-14 and Exhibit 10-2 indicate, an upward shift in the long-run trend for the high-yield country's equilibrium exchange rate should also cause the high-yield currency's real value ($q_{L/H}$) to move gradually higher as well. That is, over time, both changes in the trend for the long-run equilibrium exchange rate and the market forces driving oscillations around that long-run trend can work in concert to cause appreciation of the high-yield country's exchange rate.

What is noteworthy about Equation 10-14 is that only the interest rate spread is directly observable on the right-hand side of the equation. The trend in the long-run equilibrium exchange rate, relative inflation expectations, and relative risk premiums might be working just as hard as interest rate spreads in affecting the high-yield currency's value, but because those variables are not directly observable, nominal yield spreads are often given most, if not all, of the credit for driving changes in exchange rates and thus for the profitability of FX carry trades.

These equations also help explain why FX carry trades are often able to generate positive excess returns even as nominal yield spreads stabilize or begin to narrow. As long as the

combined impact of fiscal, economic, and monetary policy changes contributes to continued upward pressure on a high-yield currency's long-run equilibrium value, or lowers inflationary expectations in the high-yield market, or lowers a high yielder's risk premium, the net effect of these unobservable factors could lead to upward pressure on the high-yield currency's value and thereby offset the drag on the exchange rate coming from an observable decline in the yield spread. The important point to remember is that the profitability of FX carry trades sometimes depends on more than just the level and trend in the nominal yield spread.

The model also sheds light on why FX carry trades have tended to be profitable over long periods. The historical evidence suggests that the impact of nominal interest rate spreads on the exchange rate tends to be gradual. Monetary policy makers tend to adjust their official lending rates slowly over time—in part because of the uncertainty that policy makers face and in part because the authorities do not want to disrupt the financial markets. Because the monetary authorities in both the high-yield and the low-yield countries are adjusting domestic policy rates gradually over time, a high-yield country should see its short-term interest rates rise slowly relative to interest rates in the low-yield country. These slowly evolving policy rates will therefore give rise to persistence of the positive interest rate carry enjoyed by the high-yield market and in the process encourage persistence of the positive excess returns earned by FX carry trades.

Similarly, the upward trend in the equilibrium exchange rate and the downward trends in inflation expectations and risk premiums in the higher-yield market will also tend to proceed gradually. It often takes several years to determine whether structural economic changes will take root and boost the long-run competitiveness of the higher-yield country. In addition, although inflation will be favorably affected by tight monetary policies, fiscal discipline, and structural reforms, it may take several years to bring inflation expectations down to reasonably low and stable levels. Hence, all of the fundamental criteria that drive exchange rate trends over time, described in Equation 10-14, are likely to proceed gradually. Because these fundamental drivers tend to reinforce each other over time, one should expect to observe trends in real exchange rate movements and persistence in carry trade returns.

5.2.3. Equity Market Trends and Exchange Rates

Although exchange rates and equity markets sometimes exhibit positive correlation, the relationship between equity market performance and exchange rates is not stable. For instance, between 1990 and 1995, the U.S. dollar fell and the Japanese yen soared while the U.S. equity market was strong and Japanese stocks were weak. In contrast, between 1995 and early 2000, the correlation between local equity market performances and the local currency's value turned strongly positive. The U.S. dollar soared in tandem with a rising U.S. equity market, while the yen weakened in tandem with a trend decline in the Japanese equity market. *Such instability in correlation between exchange rates and equity markets makes it difficult to form judgments on possible future currency moves based solely on expected equity market performances*.

The long-run correlation between the U.S. equity market and the dollar is very close to zero, but over short- to medium-term periods, rolling correlations tend to swing from being highly positive to being highly negative, depending on market conditions. In recent years, there has been a decidedly negative correlation, with the dollar declining when the U.S. equity market was rising and vice versa. Market observers attribute this recent behavior in the dollar to the U.S. dollar's role as a safe haven asset. When investors' appetite for risk is high—that is, the market is in "risk-on" mode—investor demand for risky assets such as equities tends to rise, which drives up their prices. At the same time, investor demand for safe haven assets such

as the dollar tends to decline, which drives their values lower. The opposite has occurred when the market has been in "risk-off" mode.

EXAMPLE 10-9 Capital Flows and Exchange Rates

Monique Kwan, a currency strategist at a major foreign exchange dealer, is responsible for formulating trading strategies for the currencies of both developed market (DM) and emerging market (EM) countries. She examines two specific countries—one DM and one EM—and notes that the DM country has what is considered a low-yield safe haven currency while the EM currency has a high-yield currency whose value is more exposed to fluctuations in the global economic growth rate. Kwan is trying to form an opinion about movements in the *real* exchange rate for the EM currency (i.e., for $q_{L/H}$).

1. All else held equal, the real exchange rate for the EM currency ($q_{L/H}$) will *most likely* depreciate if:
   A. the long-run equilibrium value of the high-yield currency is revised upward.
   B. the nominal yield spread between the EM and DM countries rises over time.
   C. the expected inflation differential between the EM and DM countries is revised upward.
2. An increase in safe haven demand would *most likely*:
   A. increase the risk premium demanded by international investors to hold assets denominated in the EM currency.
   B. raise the excess return earned on carry trade strategies.
   C. exert upward pressure on the real value of the EM currency.

Kwan notes that the DM country is running a persistent current account deficit with the EM country. To isolate the influence of this chronic imbalance on exchange rates, she focuses only on the bilateral relationship between the EM and DM countries and makes the simplifying assumption that the external accounts of these two countries are otherwise balanced (i.e., there are no other current account deficits).

3. Over time, and all else held equal, the persistent current account deficit with the EM country would *most likely* lead to:
   A. a large buildup of the EM country's assets held by the DM country.
   B. an increase in the trade competitiveness of the EM country.
   C. an upward revision in the long-run equilibrium real exchange rate.

Kwan notes that because of the high yield on the EM country's bonds, international investors have recently been reallocating their portfolios more heavily toward this country's assets. As a result of these capital inflows, the EM country has been experiencing boomlike conditions. She refers to the model in Equation 10-14 to estimate the effect of these capital flows on the real exchange for the EM currency.

4. Given the current boomlike conditions in the EM economy, in the near term these capital inflows are *most likely* to lead to:
   A. a decrease in π_H^e.
   B. an increase in φ_H.
   C. an increase in $q_{L/H}$.

5. If these capital inflows led to an unwanted appreciation in the real value of its currency, the EM country's government would *most likely*:
   A. impose capital controls.
   B. decrease taxes on consumption and investment.
   C. buy its currency in the foreign exchange market.
6. If government actions were ineffective and the EM country's bubble eventually burst, this would *most likely* be reflected in an increase in:
   A. φ_H.
   B. $q_{L/H}$.
   C. $\overline{q}_{L/H}$.

Finally, Kwan turns to examining the link between the value of the DM country's currency and movements in the DM country's main stock market index. One of her research associates tells her that, in general, the correlation between equity market returns and changes in exchange rates has been found to be highly positive over time.

7. The statement made by the research associate is:
   A. correct.
   B. incorrect, because the correlation is highly negative over time.
   C. incorrect, because the correlation is not stable and tends to converge toward zero in the long run.

Solution to 1: C is correct. All else held equal, if the expected inflation differential increases, the real interest rate differential decreases, which should lead to depreciation of the real exchange rate ($q_{L/H}$). See Equation 10-13 or 10-14.

Solution to 2: A is correct. During times of intense risk aversion, investors will crowd into the safe haven currency. This tendency implies an increased risk premium demanded by investors to hold the EM currency.

Solution to 3: C is correct. Over time, the DM country will see its level of external debt rise as a result of the chronic current account imbalance. Eventually, this trend should lead to a downward revision of the DM currency's real long-run equilibrium level (via the debt sustainability channel). This is equivalent to an *increase* in the EM currency's real long-run exchange rate. Because the EM (high-yielding) currency is the base currency in the $q_{L/H}$ notation, this means an increase in $\overline{q}_{L/H}$. A is incorrect because the DM country's current account deficit is likely to lead to a buildup in DM country assets held by the EM country. B is incorrect because at some point the currency strength should contribute to deterioration in the trade competitiveness of the country with the trade surplus (the EM country).

Solution to 4: C is correct. Given the current investor enthusiasm for the EM country's assets and the boomlike conditions in the country, it is most likely that in the near term, the real exchange rate $q_{L/H}$ is increasing. At the same time, expected inflation in the EM country π^e_H is also likely increasing and—given the enthusiasm for EM assets—that the risk premium φ_H is decreasing.

Solution to 5: A is correct. To reduce unwanted appreciation of its exchange rate, the EM country would be most likely to impose capital controls to counteract the surging capital inflows. Because these inflows are often associated with overinvestment and

consumption, the EM government would not be likely to encourage these activities through lower taxes. Nor would the EM country be likely to encourage further exchange rate appreciation by intervening in the market to buy its own currency.

Solution to 6: A is correct. Episodes of surging capital flows into EM countries have often ended badly (a rapid reversal of these inflows as the bubble bursts). This is most likely to be reflected in an increase in the risk premium, φ_H. It is much less likely that a bursting bubble would be reflected in an increase in either the real exchange rate $q_{L/H}$ or its long-term equilibrium value $\overline{q}_{\mathrm{L/H}}$.

Solution to 7: C is the correct answer choice. Correlations between equity returns and exchange rates are unstable in the short term and tend toward zero in the long run.

6. MONETARY AND FISCAL POLICIES

As the foregoing discussion indicates, monetary and fiscal policies can have a significant impact on exchange rate movements. We now examine the channels through which these impacts are transmitted.

6.1. The Mundell–Fleming Model

Developed in the early 1960s, the Mundell–Fleming model remains the textbook standard for the study of monetary and fiscal policy in open economies.[18] As such, it has substantial influence on the way economists and policy makers interpret economic and financial events.

The Mundell–Fleming model describes how changes in monetary and fiscal policy affect the level of interest rates and economic activity within a country, which in turn leads to changes in the direction and magnitude of trade and capital flows and ultimately to changes in the exchange rate. As was typical of macroeconomic models in the 1960s, the model focuses only on aggregate demand. Thus, the implicit assumption is that there is sufficient slack in the economy to allow changes in output without significant price level or inflation rate adjustments. The implications of the model must be interpreted in this context.

Standard macroeconomic arguments imply that expansionary monetary and fiscal policies increase aggregate demand and output. Expansionary monetary policy affects growth, in part, by reducing interest rates and thereby increasing investment and consumption spending. Expansionary fiscal policy increases spending, either directly or via lower taxes, and thus output in the short to medium run. At the same time, expansionary fiscal policy typically exerts upward pressure on interest rates as larger budget deficits need to be financed. Based on our exchange rate model and the interest rate movements that are induced, we should expect expansionary monetary policy to lead to capital outflows and thus downward pressure on the exchange rate, while expansionary fiscal policy should lead to capital inflows and upward pressure on the exchange rate. This is indeed the key mechanism embodied in the Mundell–Fleming model.

[18]Mundell (1962, 1963) and Fleming (1962).

Although we are primarily interested in the issue of exchange rate determination and therefore the case of flexible exchange rates, it is useful to consider the implications of fixed exchange rates as well. Hence, we consider four cases involving both fixed and flexible exchange rates:

Expansionary Monetary Policy

1. *With flexible exchange rates.* Downward pressure on domestic interest rates will induce capital to flow to higher-yielding markets, putting downward pressure on the domestic currency. The more responsive capital flows are to interest rate differentials, the larger the depreciation of the currency. Depreciation of the currency will (eventually) increase net exports, reinforcing the aggregate demand impact of the expansionary monetary policy.
2. *With fixed exchange rates.* To prevent the exchange rate from depreciating, the monetary authority will have to buy its own currency in exchange for other currencies in the FX market. Doing so will tighten domestic credit conditions and offset the intended expansionary monetary policy. In the extreme case, expansionary monetary policy will be completely ineffective if the central bank is forced to fully offset the initial expansion of the money supply and allow the interest rate to rise back to its initial level. We also note that the monetary authority's ability to maintain the fixed exchange rate will be limited by its stock of foreign exchange reserves.

Expansionary Fiscal Policy

3. *With flexible exchange rates.* An expansionary fiscal policy will tend to exert upward pressure on domestic interest rates, which will in turn induce an inflow of capital from lower-yielding markets, putting upward pressure on the domestic currency. If capital flows are highly sensitive to interest rate differentials, then the domestic currency will tend to appreciate substantially. However, if capital flows are very insensitive to interest rate differentials, then the currency will tend to depreciate rather than appreciate as the policy-induced increase in aggregate demand worsens the trade balance.
4. *With fixed exchange rates.* To prevent the domestic currency from appreciating, the monetary authority will have to sell its own currency in the FX market. This expansion of the domestic money supply will reinforce the aggregate demand impact of the expansionary fiscal policy.

Despite its simplicity, the Mundell–Fleming framework provides powerful insights. First, if domestic policy makers try to (1) pursue independent monetary policies, (2) permit capital to flow freely across national borders, and (3) commit to defend fixed exchange rates, they will eventually find that the three objectives cannot be satisfied at the same time. Second, the degree of capital mobility is critical to the effectiveness of monetary and fiscal policy in an open economy. In particular, policy makers may need to impose capital controls in order to both stabilize the exchange rate and make monetary policy a viable policy instrument for domestic objectives (e.g., employment and/or price stability). This may explain why a number of emerging market economies have taken steps in recent years to impose greater control over capital flows so they can maintain their monetary policy independence on the one hand and manage their exchange rates on the other.

The specific mix of monetary and fiscal policies that a country pursues can have a profound impact on its exchange rate. Consider first the case of high capital mobility. A restrictive domestic monetary policy under floating exchange rates will give rise to an appreciation of the

EXHIBIT 10-7 Monetary–Fiscal Policy Mix and the Determination of Exchange Rates under Conditions of High Capital Mobility

	Expansionary Monetary Policy	Restrictive Monetary Policy
Expansionary Fiscal Policy	Ambiguous	Domestic Currency Appreciates
Restrictive Fiscal Policy	Domestic Currency Depreciates	Ambiguous

Source: Rosenberg (1996, 132).

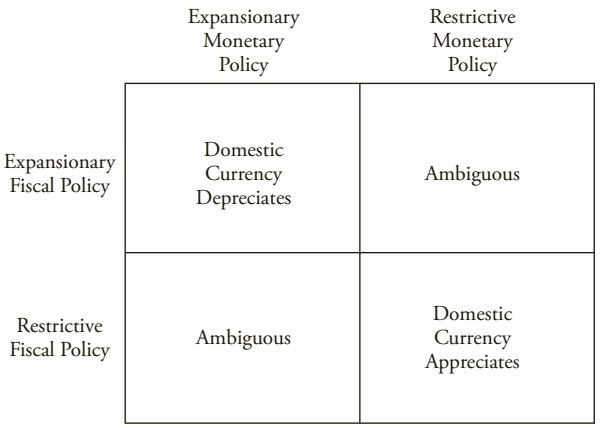
domestic currency, and an expansionary policy will give rise to currency depreciation, with a greater change in the currency when capital flows are highly mobile. On the fiscal front, an expansionary (or restrictive) fiscal policy will give rise to an appreciation (or depreciation) of the domestic currency under conditions of high capital mobility. In Exhibit 10-7, we show that the combination of an expansionary fiscal policy and a restrictive monetary policy is extremely bullish for a currency when capital mobility is high; likewise, the combination of a restrictive fiscal and an expansionary monetary policy is bearish for a currency. The effect on the currency of fiscal and monetary policies that are both expansionary or both restrictive is ambiguous under conditions of high capital mobility.

When capital mobility is low, the impact of monetary and fiscal policy changes on domestic interest rates will not induce major changes in capital flows. In such cases, monetary and fiscal policy effects on exchange rates will operate primarily through trade flows rather than capital flows. Hence, a uniformly restrictive fiscal/monetary policy mix will be bullish for a currency because this policy mix will tend to lead to an improvement in the trade balance. A uniformly expansionary fiscal/monetary policy mix will be bearish for a currency because the trade balance under such conditions would deteriorate. Combinations of expansionary fiscal and restrictive monetary policies or restrictive fiscal and expansionary monetary policies will have an ambiguous impact on aggregate demand and the trade balance, and hence on the exchange rate, under conditions of low capital mobility. Exhibit 10-8 summarizes these results.

Exhibit 10-7 is more relevant for the G-10 countries because capital mobility tends to be high in developed economies. Exhibit 10-8 is more relevant for emerging market economies that restrict the movement of capital.

A classic episode in which a dramatic shift in the policy mix caused dramatic changes in exchange rates was the case of Germany in 1990–1992. During that period, the German government pursued a highly expansionary fiscal policy to help facilitate German unification. At the same time, the Bundesbank pursued an extraordinarily tight monetary policy to combat the inflationary pressures associated with unification. The combined expansive fiscal/tight monetary policy mix drove interest rates sharply higher in Germany, and those higher interest rates were

EXHIBIT 10-8 Monetary–Fiscal Policy Mix and the Determination of Exchange Rates under Conditions of Low Capital Mobility



Source: Rosenberg (1996, 133).

then transmitted to the rest of Europe via the pegged exchange rate regime known as the European Exchange Rate Mechanism (ERM). Higher interest rates led to a marked slowdown in European growth and to a marked rise in European unemployment. Recognizing that this deterioration in the European economic climate was unsustainable, currency speculators waged an attack on the ERM regime and eventually forced several members to abandon their ERM pegs. European central banks could no longer keep interest rates high enough to defend the ERM pegs and at the same time encourage a rebound in their domestic economies.

6.2. Monetary Models of Exchange Rate Determination

In the Mundell–Fleming model, monetary policy is transmitted to the exchange rate through its impact on interest rates and output. Changes in the price level and the inflation rate play no role. Monetary models of exchange rate determination generally take the opposite perspective: Output is fixed, and monetary policy affects exchange rates first and foremost through the price level and the rate of inflation. In this section, we briefly describe two variations of the monetary approach to exchange rate determination.

6.2.1. The Monetary Approach with Flexible Prices

The monetary approach is an extension of the classical quantity theory of money to an open economy. According to the quantity theory, money supply changes are the primary determinant of price level changes. In its most extreme version, the quantity theory asserts that an X percent rise in the domestic money supply will produce an X percent rise in the domestic price level. Assuming purchasing power parity holds—that is, that changes in exchange rates reflect changes in relative inflation rates—a money supply–induced increase in domestic prices relative to foreign prices should lead to a proportional decline in the domestic currency's value, and a decrease in domestic prices should lead to an increase in currency value.

In terms of the model we developed in Section 3.3, the monetary approach focuses on the long-run equilibrium path of the nominal exchange rate. Because it assumes that PPP holds at

all times, the real exchange rate is assumed to be constant and equal to its long-run equilibrium value, $q_{L/H}$. A discrete change in the money supply will cause a proportionate change in both the equilibrium and actual nominal exchange rates. A change in the future growth rate of the money supply would change the trajectory of both the equilibrium and actual exchange rates but, in the pure monetary approach model, would not have an immediate impact on the current exchange rate.

6.2.2. The Dornbusch Overshooting Model

One of the major shortcomings of the pure monetary approach is the assumption that purchasing power parity holds at all times—that is, in both the short run and the long run. Because purchasing power parity rarely holds in either the short or the medium run, the monetary model may not provide a realistic explanation of the impact of monetary forces on the exchange rate.

To rectify that problem, Dornbusch (1976) constructed a modified monetary model of the exchange rate that assumes output prices exhibit limited flexibility in the short run but are fully flexible in the long run. The long-run flexibility of the price level ensures that any increase in the domestic money supply will give rise to a proportional increase in domestic prices and thus contribute to a depreciation of the domestic currency in the long run, which is consistent with the pure monetary model. If the domestic price level is assumed to be inflexible in the short run, however, the model implies that the exchange rate is likely to overshoot its long-run PPP path in the short run. With inflexible domestic prices in the short run, any increase in the nominal money supply results in an identical increase in the real money supply over the relevant short-term period, which in turn induces a decline in the domestic interest rate. Assuming capital is highly mobile, the decline in domestic interest rates will precipitate a capital outflow, which will cause the exchange rate to overshoot its new long-run equilibrium level to the downside on a short-run basis. Hence, in the short run, the domestic currency depreciates in both real and nominal terms. In the long run, once domestic prices rise and domestic interest rates rise in tandem, the exchange rate will recover from its overshoot level and move into line with the path predicted by the conventional monetary approach described in Section 6.2.1. The nominal exchange rate converges back to the path dictated by PPP as the domestic price level gradually adjusts and the real exchange rate returns to the equilibrium level.

6.3. The Taylor Rule and the Determination of Exchange Rates

The Mundell–Fleming and monetary models of exchange rates assume that a central bank targets, and can directly control, a country's money supply or its growth rate. Today, however, many major central banks tend to conduct policy via interest rate targets, not money supply targets. Although this practice does not invalidate the insights we have gleaned from the previous models, it is useful to consider the implications of combining a well-known interest rate targeting framework—the Taylor rule—with a basic exchange rate model.

Many central banks are charged with the responsibility of maintaining price stability and/or achieving maximum sustainable employment.[19] The key questions for individual central banks are: (1) What level of the policy rate will meet both of these policy objectives? and

[19]For example, the U.S. Federal Reserve has a dual mandate. In contrast, the Bank of Canada and the European Central Bank are explicitly charged only with price stability.

(2) How much should the policy rate rise or fall if inflation exceeds or falls short of the central bank's explicit or implicit inflation target or if the level of employment exceeds or falls short of the economy's maximum sustainable employment level?

John Taylor developed a simple mathematical rule prescribing the appropriate policy rate as a function of a central bank's neutral rate, its inflation and output targets, and observed deviations from those targets.[20] The Taylor rule is given by:

$$i = r_n + \pi + \alpha(\pi - \pi^*) + \beta(y - y^*) \tag{10-15}$$

where:

i = central bank policy rate prescribed by the Taylor rule
r_n = neutral real policy rate
π = current inflation rate
π^* = central bank's target inflation rate
y = log of the current level of output[21]
y^* = log of the economy's potential/sustainable level of output

The neutral real rate, r_n, is expected to be consistent with growth at the economy's long-run potential growth rate with stable inflation at the target rate, π^*. The neutral nominal policy rate is equal to the neutral real rate plus the target rate of inflation, π^*.[22] According to the rule, the central bank should deviate from this neutral setting only if the actual rate of inflation deviates from the targeted inflation rate (π^*) and/or the actual level of output (y) deviates from the economy's potential level of output (y^*). The magnitude of the policy rate adjustment to changes in the inflation gap ($\pi - \pi^*$) and the output gap ($y - y^*$) would be dictated by the policy response coefficients, alpha (α) and beta (β). As long as alpha and beta are both positive—Taylor proposed that alpha and beta each equal 0.5—the Taylor rule prescribes that the policy rate should rise in real terms relative to its neutral setting in response to positive inflation and output gaps (and fall in response to negative inflation and output gaps).

The Taylor rule has done a reasonably good job of explaining the trend in the U.S. federal funds rate over the Greenspan–Bernanke era, and it has done a fairly good job of explaining the trend in policy rates in other countries as well. In the United States, the actual federal funds rate has tended to move broadly in line with the recommendations of the original formulation of the Taylor rule.

Although the Taylor rule is used largely for explaining and predicting the future path of policy rates, recent research suggests that it may also provide valuable insights in determining exchange rates. To see how the Taylor rule can be used to explain the trend in exchange rates, let's first recast Equation 10-15 in real terms. This is shown in Equation 10-16:

$$r = (i - \pi) = r_n + \alpha(\pi - \pi^*) + \beta(y - y^*) \tag{10-16}$$

[20]Taylor (1993).

[21]Current and potential output are expressed as logarithms so that the difference, $(y - y^*)$, measures the percentage deviation from potential output.

[22]Inspection of Equation 10-15 shows that with this definition of the neutral nominal rate, policy will be neutral with respect to both the real and nominal interest rates only if both inflation and output are at their target levels. Various combinations of inflation and output gaps (e.g., stagflation) could imply that either the real policy rate or the nominal policy rate is at its neutral level, but not both.

Now let's assume that the central banks in two countries—say the United States and the euro area—pursue monetary policy strategies that are broadly in line with the real policy rate recommendations of the Taylor rule, described in Equation 10-16. The yield spread between U.S. and euro area real policy rates will be determined by (1) the spread between U.S. and euro area neutral real rates, (2) the spread between the actual or expected U.S.–euro area inflation gaps, (3) the spread between the actual or expected U.S.–euro areas output gaps, and (4) the relative size of the respective policy response coefficients (α and β) that the U.S. Federal Reserve and the European Central Bank normally follow.

Using the Taylor rule to substitute for the real interest rate differential in our real exchange rate model (Equation 10-12), letting the euro be the base currency, and for simplicity assuming the same policy response parameters in each market, we can write:

$$q_{USD/EUR} = \overline{q}_{USD/EUR} + (r_n^{EU} - r_n^{US}) + \alpha[(\pi_{EU} - \pi^*_{EU}) - (\pi_{US} - \pi^*_{US})] + \beta[(y_{EU} - y^*_{EU}) - (y_{US} - y^*_{US})] - (\varphi_{EU} - \varphi_{US}) \quad (10\text{-}17)$$

Equation 10-17 indicates that each of the following should strengthen the euro versus the dollar in real terms:

- An increase in the market's estimate of the euro's long-run equilibrium value.
- An increase in the ECB's estimate of the policy-neutral real interest rate relative to the Fed's estimate of the policy-neutral real rate.
- An increase in the inflation gap in the euro area relative to the inflation gap in the United States.
- An increase in the output gap in the euro area relative to the output gap in the United States.
- A decrease in the risk premium demanded for holding euro-denominated assets relative to the risk premium on dollar-denominated assets.

It is noteworthy that a rise in euro area inflation relative to the ECB's target implies appreciation of the euro, whereas a PPP framework would suggest that higher euro area inflation should be euro-negative. But in the Taylor rule framework, higher euro area inflation would compel the ECB to raise real interest rates, which would push the real interest rate differential in favor of the euro area. That, in turn, would be euro-positive. Thus bad news on euro area inflation turns out to be euro-positive, not euro-negative, as a pure monetary/purchasing power model would suggest.

6.4. Monetary Policy and Exchange Rates—The Historical Evidence

Historically, changes in monetary policy have had a profound impact on exchange rates. In the case of the U.S. dollar, the pursuit of a relatively easy monetary policy by the Federal Reserve drove U.S. real interest rates (the spread between nominal interest rates and the inflation rate) into negative territory in the late 1970s, which drove the U.S.–German real interest rate differential downward at that time. As shown earlier (Exhibit 10-5), this result contributed to a major decline in the dollar's value. The subsequent tightening of monetary policy by the Fed in the first half of the 1980s drove U.S. real interest rates up relative to real rates in Germany, which played a key role in driving the dollar higher during that period.

Japan's ultralow interest rate policy in the second half of the 1990s and into the first decade of the 2000s helped contribute to a long period of subpar performance of yen assets versus U.S. dollar assets. From the spring of 1995 until mid-2007, a 12-year period, yen money market investments dramatically underperformed U.S. dollar money market investments in U.S. dollar terms. It is widely felt that Japan's ultralow interest rate policy encouraged both Japanese and international investors to take on carry trades over much of this period, in which foreign exchange market participants borrowed heavily in yen and then invested those proceeds in higher-yielding currencies such as the Australian and New Zealand dollars.

Finally, history is replete with examples in which excessively expansionary monetary policies by central banks in emerging markets have planted the seeds of speculative attacks on their currencies. In the early 1980s, exchange rate crises in Argentina, Brazil, Chile, and Mexico were all preceded by sharp accelerations in domestic credit expansions. Emerging market policy makers appear to have learned the lessons of those past crises and policy mistakes. These policy makers have since moved toward more flexible exchange rate regimes, embraced anti-inflation or inflation-targeting policies, and built huge arsenals of foreign exchange reserves, which are now well above historical norms, as a form of insurance that could be used to defend their currencies should the need ever arise.

EXAMPLE 10-10 Monetary Policy and Exchange Rates

Monique Kwan, a currency strategist at a major foreign exchange dealer, is preparing a report on the outlook for several currencies that she follows. She begins by considering the outlook for the currency of a developed market (DM) country. This DM country has high capital mobility across its borders and a flexible exchange rate. It also has low levels of public and private debt.

Given these conditions, Kwan tries to assess the impact of each of the following policy changes using the Mundell–Fleming model and the Taylor rule.

1. For the DM currency, increasing the degree of monetary easing will *most likely*:
   A. cause the currency to appreciate.
   B. cause the currency to depreciate.
   C. have an ambiguous effect on the currency.
2. The pursuit of an expansionary domestic fiscal policy by the DM country will *most likely*:
   A. cause the domestic currency's value to appreciate.
   B. cause the domestic currency's value to depreciate.
   C. have an ambiguous effect on the domestic currency's value.

Monique Kwan's assistant prepares the following information on the DM country:

Current policy rate (nominal)	2.00%
Neutral real policy rate	2.50%
Current inflation rate	1.00%
Target inflation rate	2.00%
Current output gap	0.50%

3. Assuming that the DM central bank is following the Taylor rule and that the inflation and output gaps are equally weighted ($\alpha = \beta = 0.5$), the central bank will *most likely*:
   A. leave the policy rate unchanged.
   B. increase the policy rate by 1.00 percent.
   C. increase the policy rate by 1.25 percent.

After a period of adjustment, the current policy rate and the inflation rate are both at target levels and there is no output gap (i.e., equilibrium conditions prevail, according to the Taylor rule).

4. Given these initial equilibrium conditions, the central bank's policy response to:
   A. an expansionary fiscal policy would result in lower policy rates.
   B. an increase in growth would lead to depreciation of the DM country's currency.
   C. an increase in inflation would lead to appreciation of the DM country's currency.
5. Given these initial equilibrium conditions, the central bank's policy response to a subsequent increase in the output gap would *most likely* lead to:
   A. tighter fiscal policy.
   B. an increase in the target policy rate.
   C. a depreciation of the DM country's currency.

Next, Kwan turns her attention to an emerging market that has low levels of public and private debt. Currently, the EM country has a fixed exchange rate but no controls over international capital mobility. However, the country is considering replacing its fixed exchange rate policy with a policy based on capital controls. These proposed controls are meant to reduce international capital mobility by limiting short-term investment flows (so-called hot money) in and out of its domestic capital markets.

Kwan uses the Mundell–Fleming model to assess the likely impact of various policy changes by the EM country.

6. To maintain the exchange rate peg while increasing the degree of monetary easing, the EM country will *most likely* have to:
   A. tighten fiscal policy.
   B. decrease interest rates.
   C. buy its own currency in the FX market.
7. After replacing its currency peg with capital controls, would its exchange rate be unaffected by a tightening in monetary policy?
   A. Yes.
   B. No, the domestic currency would appreciate.
   C. No, the domestic currency would depreciate.
8. After replacing its currency peg with capital controls, the simultaneous pursuit of a tight monetary policy and a highly expansionary fiscal policy by the EM country will *most likely*:
   A. cause the currency to appreciate.
   B. cause the currency to depreciate.
   C. have an ambiguous effect on the currency.

Solution to 1: B is correct. A decrease in the policy rate would most likely cause capital to reallocate to higher-yielding investments. This would lead to currency depreciation.

Solution to 2: A is correct because an expansionary fiscal policy will lead to higher levels of government debt and interest rates, which will attract international capital flows. (In

the long run, however, an excessive buildup in debt may eventually cause depreciation pressures on the currency. This is discussed in Section 6.5.)

Solution to 3: C is correct. Under the Taylor rule, the prescribed central bank policy rate is equal to:

$$i = 2.50\% + 1.00\% + \frac{1}{2}(1.00\% - 2.00\%) + \frac{1}{2}(0.50\%) = 3.25\%$$

This requires a 1.25 percent increase from the current 2.00 percent policy rate.

Solution to 4: C is correct. Above-target inflation will lead to monetary tightening, which should lead to currency appreciation. A is incorrect because an expansionary fiscal policy is likely to both lead to economic growth moving above potential and also add to inflationary pressures. Both effects should lead to an increase in the central bank's policy rate. B is incorrect because an increase in economic growth above potential, all else held equal, will lead to an increase in the policy rate determined by the Taylor rule. A higher policy rate will likely lead to currency appreciation.

Solution to 5: B is correct. An increase in the output gap ($y - y^*$) means the economy is growing above potential. Under the Taylor rule, the central bank will tighten monetary policy. A is incorrect because the central bank does not control fiscal policy, only monetary policy. C is incorrect because an increase in the output gap will lead to a tightening of monetary policy, which, all else being equal, should lead to currency appreciation, not depreciation.

Solution to 6: C is correct. The looser monetary policy will lead to exchange rate depreciation. To counter this effect and maintain the currency peg, the central bank will have to intervene in the FX market, buying its own currency. A is incorrect because tighter fiscal policy is associated with lower interest rates and is therefore likely to increase rather than mitigate the downward pressure on the domestic currency. Similarly, B is incorrect because a move to lower interest rates would exacerbate the downward pressure on the currency and hence the pressure on the peg.

Solution to 7: B is correct. In general, capital controls will not completely eliminate capital flows but will limit their magnitude and responsiveness to investment incentives such as interest rate differentials. At a minimum, flows directly related to financing international trade will typically be allowed. The exchange rate will still respond to monetary policy. With limited capital mobility, however, monetary policy's main influence is likely to come through the impact on aggregate demand and the trade balance. A tighter domestic monetary policy will most likely lead to higher interest rates and less domestic demand, including less demand for imported goods. With fewer imports, and exports held constant, there will be modest upward pressure on the currency.

Solution to 8: C is correct because (1) capital mobility is low, so the induced increase in interest rates is likely to exert only weak upward pressure on the currency; (2) the combined impact on aggregate demand is ambiguous; and (3) if aggregate demand increases, the downward pressure on the currency due to a worsening trade balance may or may not fully offset the upward pressure exerted by capital flows.

6.5. Fiscal Policy and the Determination of Exchange Rates

Virtually all of the exchange rate models that economists have devised agree that the pursuit of relatively easy domestic monetary policies will tend to exert downward pressure on a domestic currency's value, while the pursuit of relatively tight domestic monetary policies will tend to exert upward pressure on a domestic currency's value. In this section, we shift our focus to examine what role fiscal policy changes have in determining exchange rates.

Despite common agreement on the role of monetary policy in determining exchange rates, *fiscal policy's impact on exchange rates is ambiguous* because fiscal impulses are transmitted to exchange rates through a variety of channels, some of which transmit positive influences on a currency's value while others transmit negative influences. Whether a given change in fiscal policy will result in an increase or a decrease in a currency's value will depend on whether the positive channels dominate the negative ones or vice versa.

The Mundell–Fleming model is essentially a short-run model of exchange rate determination. It makes no allowance for the long-term effects of budgetary imbalances that typically arise from sustained fiscal policy actions. The portfolio balance approach to exchange rate determination remedies this.

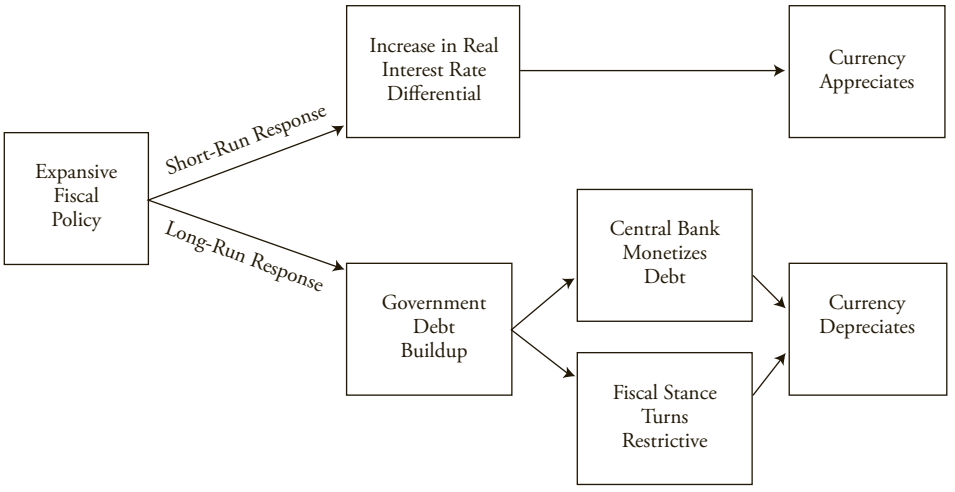
In the **portfolio balance approach**, global investors are assumed to hold a diversified portfolio of domestic and foreign assets, including bonds. Their desired allocation is assumed to vary in response to changes in expected return and risk considerations. In this framework, a steady increase in the supply of domestic bonds outstanding, generated by a continued widening of the government budget deficit, will be willingly held only if asset holders are compensated in the form of a higher expected return. Such a return could come from (1) higher interest rates and/or higher risk premiums, (2) immediate *depreciation of the currency to a level sufficient to generate anticipation of gains from subsequent currency appreciation*, or (3) some combination of the two. The second mechanism, currency adjustments required to achieve or maintain global asset market equilibrium, is the crux of the portfolio balance approach.

One of the major insights one should draw from the portfolio balance model is that *in the long run, governments that run large budget deficits on a sustained basis could eventually see their currencies decline in value*.

The Mundell–Fleming and portfolio balance models can be combined into a single integrated framework in which expansionary fiscal policy under conditions of high capital mobility may be positive for a currency in the short run but negative in the long run. Exhibit 10-9 illustrates this concept. A domestic currency may rise in value when the stimulative fiscal policy is first put into place. As deficits mount over time and the government's debt obligations rise, however, the market will begin to wonder how that debt will be financed. If the volume of debt rises to levels that are believed to be unsustainable, the market may believe that pressure will eventually have to be brought to bear on the central bank to monetize the debt—that is, for the central bank to buy the government's debt with newly created money. Such a scenario would clearly lead to a rapid reversal of the initial currency appreciation.

Alternatively, the market may believe that the government's fiscal stance will eventually have to shift toward significant restraint to restore longer-run sustainable balance to its fiscal position. A reversal of the fiscal stance that initially drove the currency higher would set forces in motion for a reversal of the initial currency appreciation.

EXHIBIT 10-9 The Short-Run and Long-Run Responses of Exchange Rates to Changes in Fiscal Policy



Source: Rosenberg (2003).

EXAMPLE 10-11 Fiscal Policy and Exchange Rates

Monique Kwan is continuing her analysis of the foreign exchange rate outlook for selected countries. She examines a DM country that has a high degree of capital mobility and a floating-rate currency regime. Kwan notices that although the current outstanding stock of government debt is low as a percentage of GDP, it is rising sharply as a result of expansionary fiscal policy. Moreover, projections for the government debt-to-GDP ratio point to further increases well into the future.

Kwan uses the Mundell–Fleming and portfolio balance models to form an opinion about both the short-run and long-run implications for the DM country's exchange rate.

1. Over the short run, Kwan is *most likely* to expect:
   A. appreciation of the DM's currency.
   B. an increase in the DM's asset prices.
   C. a decrease in the DM's risk premium.
2. Over the medium term, as the DM country's government debt becomes harder to finance, Kwan would be *most likely* to expect that:
   A. fiscal policy will turn more accommodative.
   B. the mark-to-market value of the debt will increase.
   C. monetary policy will become more accommodative.

3. Assuming that the DM country's government debt becomes harder to finance and there is no change in monetary policy, Kwan is *most likely* to expect that over the longer term, there will be a fiscal policy response that would lead to:
   A. currency appreciation as yields rise.
   B. currency depreciation as yields decline.
   C. an ambiguous impact on the currency, depending on which effect prevails.

Solution to 1: A is correct. The DM country currently has a low debt load (as a percentage of GDP), and in the short run, its expansionary fiscal policy will lead to higher interest rates and higher real rates relative to other countries. This path should lead to currency appreciation. The higher domestic interest rates will (all else held equal) depress local asset prices (so B is incorrect), and the rising debt load is likely to increase rather than decrease the risk premium (so C is incorrect).

Solution to 2: C is correct. This is because as government debt becomes harder to finance, the government will be tempted to monetize the debt through an accommodative monetary policy. A is incorrect because an inability to finance the debt will make it hard for fiscal policy to become more accommodative. B is incorrect because as investors demand a higher risk premium (a higher return) for holding the DM country's debt, the mark-to-market value of the debt will decline (i.e., bond prices will decline and bond yields will increase).

Solution to 3: B is correct. As the DM country's debt ratios deteriorate, foreign investors will demand a higher rate of return to compensate them for the increased risks. Assuming that the central bank will not accommodate (monetize) the rising government debt, the most likely fiscal response is an eventual move toward fiscal consolidation—reducing the public deficit and debt levels that were causing the debt metrics to deteriorate. This policy adjustment would involve issuing fewer government bonds. All else being equal, bond yields would decrease, leading to a weaker domestic currency.

A is not the most likely answer because currency appreciation is not likely to accompany rising yields when the government is having difficulty financing its deficit. There would be a rising risk premium (a deteriorating investor appetite) for holding DM assets and hence a currency appreciation would be unlikely despite high DM yields. To avoid paying these high yields on its debt, the DM government would eventually have to take measures to reduce its deficit spending. This approach would eventually help reduce investor risk aversion and DM yields. C is incorrect because given the deterioration in the DM's debt metrics, a depreciation of its exchange rate is likely to be an important part of the restoration of financial market equilibrium.

7. EXCHANGE RATE MANAGEMENT: INTERVENTION AND CONTROLS

Capital flow surges can be both a blessing and a curse. Capital inflows can be a blessing if they enable growing economies to bridge the gap between domestic investment and domestic savings. They can be a curse, however, if they fuel boomlike conditions, asset price bubbles,

and an overshooting of the currency into overvalued territory. Problems arise when inflows of short-term capital eventually reverse, because the resulting outflow of capital can trigger a major economic downturn, a significant decline in asset prices, and a major depreciation of the currency.

Capital flow surges planted the seeds of three major currency crises in the 1990s—the ERM crisis in the fall of 1992, the Mexican peso crisis in late 1994, and the Asian currency and financial crisis in 1997–1998. Each crisis episode was preceded by a surge in capital inflows that led to a buildup of huge, highly leveraged speculative positions by local as well as international investors in currencies that eventually came under heavy speculative attack. In the run-up to the ERM crisis, investors—believing that European yield convergence would occur as European monetary union approached—took on highly leveraged long positions in the higher-yielding European currencies financed by short positions in the lower-yielding European currencies. Likewise, in the run-up to the Mexican peso crisis, investors and banks were highly leveraged and made extensive use of derivative products in taking on speculative long Mexican peso/short U.S. dollar positions. And in the run-up to the Asian financial crisis, Asian companies and banks were highly leveraged as they took on a huge volume of short-term dollar- and yen-denominated debt to fund local activities. In each case, the sudden unwinding of those leveraged long speculative positions triggered the attacks on the currencies.

Surges in capital inflows often are driven by a combination of *pull* and *push* factors, with both factors capable of generating bubblelike conditions. Pull factors represent a favorable set of developments that encourage overseas capital to flow toward a particular country. These factors include:

- Better economic management by policy makers.
- Expected declines in inflation and inflation volatility.
- More flexible exchange rate regimes.
- Improved current account balances.
- Declines in public-sector and private-sector debt burdens.
- A significant buildup in FX reserves, which can be used as a buffer against future speculative attacks.
- Privatization of state-owned entities.
- Liberalization of financial markets.
- Lifting of foreign exchange regulations and controls.
- Strong and sustained economic growth, which works to enhance the expected long-run return on real and financial assets.
- Improved fiscal positions.
- Sovereign ratings upgrades.

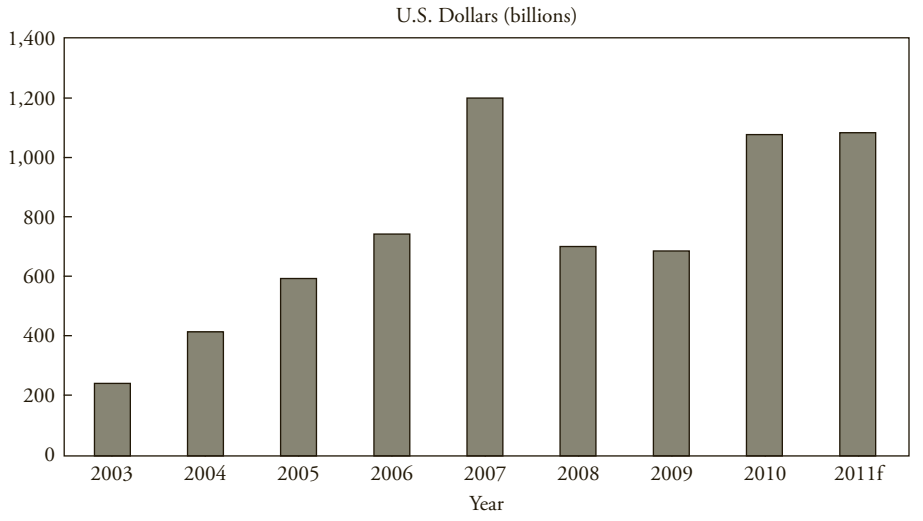
Push factors represent a favorable set of factors that emanate not from the recipient country per se but rather from the primary sources of internationally mobile capital, notably the investor base in industrial countries. The pursuit of low interest rate policies in industrial countries has often encouraged investors in those markets to move funds offshore to earn higher returns than can be earned domestically. Japan's ultralow interest rates encouraged investors to move funds into higher-yielding markets in Australia, New Zealand, and elsewhere in Asia. The pursuit of a low interest rate policy by the Federal Reserve encouraged U.S. investors to allocate more of their funds to emerging markets in the mid and late 2000s. Another important push factor is the long-run trend in asset allocation by industrial country investors. For example, U.S. fund managers have traditionally had underweight exposures to

emerging market assets, but with the weight of emerging market equities in broad global equity market indexes on the rise, U.S. investor allocation to EM equities is likely to rise in lockstep. Because the EM share of world GDP is now around 40 percent, up from 17 percent in the 1960s, the share of funds allocated to emerging markets is likely to increase on a trend basis. Notably, just a one percentage point annual increase in industrial country investor allocations to emerging market assets could add roughly $350 billion to $500 billion to the annual flow of net private capital to emerging markets. That amount would be a sizable chunk, considering that net private capital flows to emerging markets in 2010 totaled roughly $900 billion.

As indicated in Exhibit 10-10, net private capital inflows to emerging markets rose steadily between 2003 and 2007, posting nearly a sixfold increase over the period. Both push and pull factors contributed to that surge in capital flows. Net private capital flows to emerging markets tumbled in 2008 and 2009 as heightened risk aversion during the global financial crisis prompted investors to unwind some of their EM exposures in favor of U.S. assets. Since 2009, capital flows to emerging markets have once again started to rise. Emerging market investments appear to be gaining favor again in part because most EM economies appear to have weathered the global financial crisis better than most industrial economies, while the pursuit of ultralow interest rate policies in the United States, euro area, and Japan has encouraged global investors to invest in higher-yielding EM assets.

The key question facing EM policy makers is how best to respond to the latest wave of capital inflows. As discussed earlier, some of the major crisis episodes of the 1990s were instigated, in part, by excessive surges in capital inflows that eventually were reversed. This time around, policy makers are intent on ensuring that history does not repeat itself. Hence, policy makers appear to be making a concerted effort to either (1) resist or deflect such inflows through the use of capital controls or (2) prevent capital inflows from pushing EM currency values to overvalued levels by intervening more heavily in the foreign exchange market.

EXHIBIT 10-10 Net Private Capital Flows to Emerging Markets



Source: Adapted from Suttle et al. (2011, 1–2).

The International Monetary Fund offers the following guidelines to policy makers who face an unwanted surge in capital inflows: Assuming that the EM currency is undervalued, the appropriate policy response would be to allow the currency to appreciate. The danger, of course, is that the appreciation ends up being excessive, leading eventually to overvaluation and a loss of competitiveness. If the exchange rate is already fairly valued or overvalued, then the appropriate policy step would be to intervene. If there is no inflation threat, the authorities could engage in **unsterilized intervention**, which would expand the monetary base and encourage short-term interest rates to move lower. Lower interest rates, in turn, might help discourage future capital inflows. If, however, inflation is a concern, then this intervention would need to be sterilized. In a **sterilized intervention** operation, EM authorities would sell domestic securities to the private sector to mop up any excess liquidity created by its FX intervention activities. The end result would be that the monetary base and the level of short-term interest rates would not be altered by the intervention operation. If there are limits on the capacity of EM countries to engage in sterilized intervention operations, then a tightening in fiscal policy might be needed to help slow domestic demand and help alleviate some of the upward pressure on the EM currency.

If all of the preceding failed to stop the capital flow–induced upward pressure on the EM currency, then capital controls might have to be considered as a final line of resistance in preventing capital flows from pushing currency values and asset prices to undesirable levels. At one time, capital controls were frowned on as a policy tool for curbing undesired surges in capital inflows. It was generally felt that such controls tended to generate distortions in global trade and finance and that, in all likelihood, market participants would eventually find ways to circumvent the controls, which, in turn, would require an ever more pervasive set of regulations and controls to combat the evasion of the initial set of controls. It has also been felt that the use of capital controls to resist upward pressure on real exchange rates in countries running high current account surpluses might exacerbate global imbalances. Furthermore, many observers feared that capital controls imposed by one country could deflect capital flows to other countries, which could complicate monetary and exchange rate policies in those economies. Despite such concerns, the IMF believes that the benefits associated with capital controls may exceed the associated costs. Hence, the IMF now considers capital controls a legitimate part of EM policy makers' tool kit. Given the painful lessons that EM policy makers have learned from previous episodes of capital flow surges, it is now believed that under certain circumstances, capital controls may be needed to prevent exchange rates from overshooting, asset bubbles from forming, and future financial conditions from deteriorating.

Although a case can be made for central bank intervention and capital controls to limit the potential damage associated with unfettered inflows of overseas capital, the key issue for policy makers is whether intervention and capital controls will actually work in terms of (1) preventing currencies from appreciating too strongly, (2) reducing the aggregate volume of capital inflows, and (3) enabling monetary authorities to pursue independent monetary policies without having to worry about whether changes in policy rates might attract too much capital from overseas.

Evidence on the effectiveness of central bank intervention suggests that, in the case of industrial countries, the volume of intervention is often quite small relative to the average daily turnover of G-10 currencies in the foreign exchange market. Hence, most studies conclude that the effect of intervention in developed market economies is either statistically insignificant or quantitatively unimportant. For most developed market countries, the ratio of official FX reserves held by the respective central banks to the average daily turnover of foreign exchange trading in that currency is negligible. With insufficient resources in most industrial countries'

reserve arsenals, there simply is not enough firepower available to significantly affect the supply of and demand for foreign exchange.[23]

The evidence on the effectiveness of central bank intervention in emerging market currencies is more mixed. Intervention appears to contribute to lower EM exchange rate volatility, but no statistically significant relationship has emerged between the level of EM exchange rates and intervention. Some studies find, however, that EM policy makers might have greater success than their industrial country counterparts in terms of their ability to influence the level and path of their exchange rates, because the ratio of EM central bank FX reserve holdings to average daily FX turnover in their domestic currencies is actually quite sizable. With considerably greater firepower in their reserve arsenals, emerging market central banks appear to be in a stronger position than their developed market counterparts to influence the level and path of their exchange rates. What's more, with emerging market central banks' FX reserve holdings expanding at a near-record clip in the past decade, the effectiveness of sterilized intervention may be stronger now than in the past.

The evidence on the effectiveness of capital controls in terms of resisting or deflecting capital inflows is mixed. In a recent survey of empirical studies on the effectiveness of capital controls in emerging markets by Magud, Reinhart, and Rogoff (2011), the authors optimistically concluded that capital controls on inflows (1) make monetary policy more independent, (2) can alter the composition of capital flows, and (3) can reduce real exchange rate pressures in certain instances (although the evidence on this point is more controversial). The study also found, however, that "capital controls on inflows seem not to reduce the volume of net flows." To a large extent, *the relative success or lack thereof of controls on capital inflows hinges on the magnitude and persistence of the inflows that policy makers are seeking to resist. The more persistent those flows are and the larger their magnitude, the less likely it is that capital controls will be effective* in stemming upward pressure on the real exchange rate.

EXAMPLE 10-12 Exchange Rate Management: Intervention and Controls

Monique Kwan now turns her attention to an EM country that is experiencing a surge of capital inflows. This country recently improved its fiscal position through the privatization of state-owned assets and has seen its sovereign credit rating upgraded (both pull factors). Although the country's current policies allow a high degree of capital mobility, it is becoming concerned about the potential future impact of these capital inflows.

1. The EM country is *more likely* to engage in sterilized intervention if:
   A. its inflation rate is high.
   B. its currency is undervalued.
   C. its currency appreciation is caused by push factors rather than pull factors.

[23]If a central bank is intervening in an effort to weaken, rather than strengthen, its own currency, it could (at least in principle) create and sell an unlimited amount of its currency and accumulate a correspondingly large quantity of FX reserves. As discussed earlier, however, persistent intervention in the FX market undermines the efficacy of monetary policy for domestic purposes.

2. If the EM country used capital controls instead, this approach would:
   A. lead to a less independent monetary policy.
   B. be more likely to succeed when capital flows are less persistent.
   C. require that the ratio of foreign exchange reserves to FX market turnover be high.

Solution to 1: A is correct. A country would likely choose sterilized intervention if it had a high inflation rate, because unsterilized intervention would add to the monetary base and possibly increase inflationary pressures. B is incorrect because an undervalued currency is likely to lessen the probability of intervention, sterilized or unsterilized, in response to capital inflows. C is incorrect because both push and pull factors can lead to bubble conditions, excessive exchange rate appreciation, and the unwanted buildup of excessive speculative positions.

Solution to 2: B is correct. Although the empirical evidence is mixed, to a large extent the relative success of capital controls depends on the magnitude and persistence of the capital inflows that the policy makers are trying to resist (the less persistent, the higher the effectiveness). A successful capital control policy tends to increase, not decrease, monetary policy independence. As a regulatory measure, capital controls do not depend directly on the level of foreign exchange reserves. (This is in contrast to FX intervention, for which the likelihood of success would be enhanced by having relatively large foreign reserves.)

8. CURRENCY CRISES

History is replete with examples of currencies that have come under heavy selling pressure within short windows of time. These episodes often occurred suddenly, with many (if not most) investors surprised by the timing of the crisis.

If market participants correctly anticipated the onset of a currency crisis, one would expect to see a substantial widening in interest rate differentials to reflect the higher likelihood of a major currency depreciation as the impending crisis drew nearer. In addition, one would expect to see expectations of impending currency weakness built into the consensus forecasts of exchange rates by international economists. Finally, because many currency crisis episodes have coincided with banking and other forms of financial crises, one would expect that credit agency risk assessments would be signaling an impending crisis.

Unfortunately, a careful reading of the history of currency and financial crisis episodes suggests that this has not been the case. For example, interest rate differentials failed to widen measurably in anticipation of the ERM crisis of 1992–1993, the Mexican peso crisis of 1994–1995, or the Asian financial crisis of 1997–1998. More recently, risk premiums embedded in many spread products in the U.S. financial markets failed to widen in advance of the 2007–2009 financial crisis. Particularly noteworthy is the fact that neither the consensus forecasts of private and government economists nor the risk assessments by credit rating agencies provided adequate warning of the impending financial shocks and the ensuing economic downturns.

If anything, the evidence seems more consistent with the view that market participants are taken by surprise when a crisis occurs. Indeed, more often than not, investors and

borrowers tend to be leaning the wrong way in terms of their portfolio positioning at the time of an attack. Once a wave of selling begins, investors and borrowers must immediately reposition their portfolios to avoid excessive capital losses. Such repositioning often works to intensify selling pressure on the currency. It is this massive liquidation of vulnerable positions, often reinforced by speculative selling, that is largely responsible for the seemingly excessive exchange rate movements that typically occur during a currency crises.

Because most crisis episodes have not been adequately anticipated, economists at the IMF and at leading investment banks and think tanks have spent a great deal of time in recent years developing early warning systems to help policy makers and investors better position themselves the next time an impending crisis draws nearer. One of the problems in developing an early warning system is that views on the underlying causes of currency crises differ greatly. One school of thought contends that currency crises tend to be precipitated by deteriorating economic fundamentals, while a second school contends that currency crises can occur out of the blue, with no evidence of deteriorating fundamentals preceding them.

According to the first school of thought, evidence of a secular deterioration in economic fundamentals often precedes most crisis episodes. If currency crises tend to be preceded by weak economic fundamentals, and the trend in those economic fundamentals deteriorates steadily and predictably, then it should be possible to construct an early warning system to anticipate when a currency might be vulnerable.

The second school of thought argues that although evidence of deteriorating economic fundamentals might explain a relatively large number of currency collapses, there might be cases in which economies with relatively sound fundamentals could see their currencies come under attack because of (1) a sudden adverse shift in market sentiment totally unrelated to economic fundamentals or (2) contagion or spillover effects arising from crisis developments in other markets. If that is the case, then it might not be possible to construct a reliable early warning system to predict future currency crises.

Recognizing that no single model can correctly anticipate the onset of all crisis episodes, an early warning system might nevertheless be useful in assisting (1) fund managers in structuring their global investment portfolios and (2) policy makers in taking the necessary steps to avoid the impending crisis in the first place.

An ideal early warning system would need to incorporate a number of important features. First, it should have a strong record not only in terms of predicting actual crises but also in terms of avoiding the frequent issuance of false signals. Second, it should include macroeconomic indicators whose data are available on a timely basis. If data arrive with a long lag, one runs the risk that a crisis could be underway before the early warning system starts flashing red. Indeed, an ideal early warning system would be one that starts flashing red well in advance of an actual currency crisis to allow market participants sufficient time to adjust or hedge their portfolios before the crisis hits. Third, because currency crises tend to be triggered in countries with a number of economic problems, not just one, an ideal early warning system should be broad-based, incorporating a wide range of symptoms that crisis-prone currencies might exhibit.

The IMF conducted a study analyzing the behavior of 10 key macroeconomic variables around the time of currency crises in 50 countries over the 1975–1997 period. Although the behavior of these variables often differed from one crisis to another, a number of stylized facts emerged from the study:

- In the period leading up to a crisis, the real exchange rate is substantially higher than its mean level during tranquil periods.

- Somewhat surprisingly, the trade balance displays no significant difference between its behavior in precrisis periods and in tranquil periods.
- Foreign exchange reserves tend to decline precipitously as the crisis approaches.
- On average, there is some deterioration in the terms of trade in the months leading up to a crisis.
- Inflation tends to be significantly higher in precrisis periods compared with tranquil periods.
- The ratio of M2, a measure of money supply, to bank reserves tends to rise in the 24-month period leading up to a crisis and then plummets sharply in the months immediately following a crisis.
- Broad money growth in nominal and real terms tends to rise sharply in the two years leading up to a currency crisis, peaking around 18 months before a crisis hits.
- Nominal private credit growth also tends to rise sharply in the period leading up to a crisis.
- Currency crises are often preceded by a boom–bust cycle in financial asset (equity) prices.
- Real economic activity displays no distinctive pattern ahead of a crisis but falls sharply in the aftermath.

Given the stylized behavior of these key economic variables prior to the onset of many major currency crises, model builders at the IMF and at leading investment banks have made numerous attempts to construct *composite* early warning systems that incorporate a number of key economic variables into a single index of crisis vulnerability. Although no single index can capture all of the economic and financial developments leading up to each and every crisis, such indexes may nevertheless prove useful in assessing potential negative tail risks that might be lurking around the corner.

EXAMPLE 10-13 Currency Crises

Monique Kwan now turns her attention to the likelihood of crises in various emerging market currencies. She discusses this matter with a research associate, who tells her that the historical record of currency crises shows that most of these episodes were not very well anticipated by investors (in terms of their positioning), by the bond markets (in terms of yield spreads between countries), or by major credit rating agencies and economists (in terms of the sovereign credit ratings or forecasts, respectively).

1. The research associate is *most likely*:
   A. correct.
   B. incorrect, because most credit rating agencies and economists typically change their forecasts prior to a crisis.
   C. incorrect, because investor positioning and international yield differentials typically shift prior to a crisis.

Kwan delves further into the historical record of currency crises. She concludes that even countries with relatively sound economic fundamentals can fall victim to these crisis episodes and that these attacks can occur when sentiment shifts for reasons unrelated to economic fundamentals.

2. Kwan's conclusion is *most likely*:
   A. correct.
   B. incorrect, because there are few historical crises involving currencies of countries with sound economic fundamentals.
   C. incorrect, because there are few historical episodes in which a sudden adverse shift in market sentiment occurs that is unrelated to economic fundamentals.

To better advise the firm's clients on the likelihood of currency crises, Kwan tries to formulate an early warning system for these episodes. She recognizes that a typical currency crisis tends to be triggered by a number of economic problems, not just one.

3. Kwan's early warning system is *least likely* to indicate that there might be an impending crisis when there is:
   A. an expansionary monetary policy.
   B. an overly appreciated real exchange rate.
   C. rising foreign exchange reserves at the central bank.
4. Kwan's early warning system would *most likely* be designed to:
   A. have a strong record of predicting actual crises, even if it generates a lot of false signals.
   B. be based on a wide variety of economic indicators, including those for which data are available only with a significant lag.
   C. start flashing well in advance of an actual currency crisis to give market participants time to adjust or hedge their portfolios before the crisis hits.

Solution to 1: A is correct. Currency crises often catch most market participants and analysts by surprise.

Solution to 2: A is correct. Even countries with sound economic fundamentals can be subject to a currency crisis, including instances when market sentiment shifts for noneconomic reasons.

Solution to 3: C is correct. A high level of foreign exchange reserves held by a country typically decreases the likelihood of a currency crisis.

Solution to 4: C is correct. Early warnings are a positive factor in judging the effectiveness of the system, whereas false signals and the use of lagged data would be considered negative factors.

9. SHORTER-TERM FORECASTING TOOLS

Correctly predicting the direction and magnitude of exchange rate movements on a more or less consistent basis is not an easy task. Economists have developed a range of theories to explain how exchange rates are determined, but the overwhelming body of evidence from hundreds of empirical studies indicates that fundamentals-based models, although useful in explaining the longer-term trends in exchange rates, have not had much success in explaining short- and medium-term trends. Indeed, a random walk characterizes exchange rate movements better than most conventional fundamentals-based exchange rate models in the short run.

Frustrated with the performance of such models, fund managers have turned to other types of forecasting tools to help them forecast the direction of exchange rates. Because the performance of fund managers is often evaluated over relatively short time spans, investor attention tends to focus on short-run forecasting tools such as technical analysis and order flow, sentiment, and positioning indicators for a better reading on the short-run pressures driving exchange rates. Although technical analysis has been widely used in the FX arena for decades, investor interest in order flow, sentiment, and positioning has been more recent. This newly found interest stems in part from the fact that data on order flow and positioning have only recently become more readily available.

9.1. Technical Analysis

Followers of technical analysis believe that the market tips its hand ahead of time as to the likely future direction of exchange rates. That is, if the upside and downside moves in exchange rates exhibit a tendency to recur in a systematic manner, then the study of past price action can allow investors to draw conclusions regarding the likely direction and magnitude of future price movements.

Technical trading rules come in many different forms. Some are designed to identify market trends or market reversals, while others are designed to identify:

- Overbought or oversold conditions
- Relative strength
- Support and resistance levels.

A recent survey found that approximately 90 percent of FX traders used some technical analysis input to help position themselves on an intraday, daily, or weekly basis.

The question, of course, is whether technical analysis represents a reliable approach to taking positions in the FX market. Numerous academic studies of the FX market conducted in the 1970s, 1980s, and early 1990s overwhelmingly concluded that a variety of trend-following trading rules—such as moving average crossover trading rules and filter rules—would have generated significant profits had such models been actively followed over those periods. But updated studies for the 1995–2010 period indicate that trend-following trading rules have not fared as well as they did in earlier periods. Indeed, those updated studies now find that whatever profitability previously existed for technical trading rules has vanished during the past 15 years.

The principal reason why trend-following trading rules have become less profitable since the mid-1990s is that persistent and pronounced exchange rate swings are not occurring as often as they did in the past. Most of the profits over the 1973–1995 period came from participating in a few correctly predicted and very large exchange rate moves. Because such large swings have become less frequent, the positive excess returns that could be earned by adhering to the dictates of trend-following rules have largely been eliminated.

A 2008 study by Pukthuanthong-Le and Thomas documents the decline in returns that trend-following trading rules offer. According to their findings, "the era of easy profits from simple trend-following strategies in major currencies is over." Exhibit 10-11 presents their findings for six key currencies versus the U.S. dollar, with returns broken down into six subperiods: 1975–1979, 1980–1984, 1985–1989, 1990–1994, 1995–1999, and 2000–2006. As the exhibit shows, average annual excess returns were highly positive in the 1970s and 1980s but began to drop off in the 1990s and declined more substantially in the early 2000s.

EXHIBIT 10-11 Liquid Currencies' Mean Profit Performance Using Trend-Trading Moving Average Rules, 1975–2006

Currency	1975–1979	1980–1984	1985–1989	1990–1994	1995–1999	2000–2006
Japanese yen	11.10	4.81	11.47	5.82	10.04	−2.29
German mark/euro	6.81	6.58	9.77	0.79	10.37	2.20
British pound	11.03	7.80	3.31	1.21	−5.90	−0.17
Swiss franc	6.70	6.31	7.98	2.55	−0.51	−0.22
Canadian dollar	3.52	0.89	1.58	1.22	−0.96	0.11
Australian dollar	–	–	−0.78	−0.09	−1.02	−1.44
Portfolio	7.83	5.28	5.66	1.92	2.00	−0.30

Source: Pukthuanthong-Le and Thomas (2008).

Pukthuanthong-Le and Thomas do, however, offer a glimmer of hope for technically based traders. The authors suggest that "trending might be a feature confined to currencies in the early years of a floating rate regime." They find that significant positive excess returns could have been earned from 2002 to 2006 if FX traders applied moving average trading rules to selected emerging market currencies, such as the Brazilian real, Mexican peso, South African rand, and Russian ruble, although the peso market might have matured and may no longer be as profitable.

One reason for the success of technically based trading in the emerging market currencies might be that the EM currency market is less crowded at present, similar to the G-10 market 30 years ago. A less crowded market implies that profitable trading opportunities might not have been completely arbitraged away.

Another reason for the success of technically based trading in emerging market currencies is that many such currencies have experienced significant real appreciation during the 2000s. Although it might not have been possible to observe and quantify all of the favorable structural and policy-related factors taking place in Asia, Latin America, and Eastern Europe/Middle East/Africa in real time, the net positive impact of these changes on currency values could have been indirectly captured by simple trend-following trading rules.

Interestingly, a case can be made for monitoring technical price action in the G-10 currencies as well, even though such activities are unlikely to add incremental return to a currency portfolio. A number of studies find that technical analysis may be useful for controlling risk, even if technically based strategies fail to boost average returns. A novel experiment conducted by Riccardo Curcio and Charles Goodhart (1992) of the London School of Economics involved separating their students into two groups to see which group could generate the best FX trading performance. One group formulated currency trading strategies with the assistance of a technically based forecasting service, while the other group did not. The results of the experiment indicated that the total return earned by both groups was roughly the same. That is, the use of a technical model input did not contribute to above-average total return performance. Interestingly, however, the standard deviation of returns was considerably smaller for the group using the technical model input than was the case for the group that received no assistance at all.

Curcio and Goodhart noted that without the assistance of technical analysis, students tended to follow extreme contrarian strategies that magnified the volatility of their total

returns. The group of students who closely followed the dictates of a technically based trading system were able to avoid taking extreme contrarian positions. Hence, Curcio and Goodhart concluded that technically based trading systems might be useful for controlling risk, even if those strategies failed to enhance portfolio returns.

Curcio and Goodhart's findings have applicability for FX fund managers who implement carry trade strategies. Although carry trade strategies have been found to generate attractive returns over long periods, the distribution of those returns suggest that carry trades are prone to significant downside tail risk from time to time during carry trade unwinds. Therefore, a trend-following trading system overlaid on a carry trade strategy could warn investors to step aside when such unwinds are occurring.

9.2. Order Flow, Sentiment, and Positioning

If price trends are of limited use in predicting future exchange rate movements, what about other market indicators? Most studies find a positive, *contemporaneous* correlation between the trends in order flow, sentiment, and positioning indicators on the one hand and the trend in exchange rates on the other. The evidence is more mixed on whether these indicators have any value in terms of *predicting* the future direction of exchange rates on a short- to medium-term basis.

9.2.1. FX Dealer Order Flow

In recent years, market participants have become increasingly interested in FX dealer customer flow data. One of the characteristics that distinguishes the FX market from the world equity market is that the FX market has considerably less transparency. Equity market disclosure requirements mandate that all trades be posted instantly. Thus, volume and price data are instantly available to all parties. Not so in the FX market—no such disclosure requirements exist, which means that order flow information is not immediately available to all parties. Thus, a large FX dealer's order book could be of value to investors if the order flow data were shown to have predictive value on a short- to medium-term basis.

Most studies find a strong, positive, *contemporaneous* correlation between cumulative customer order flow and exchange rates over short periods of time (i.e., on an intraday or daily basis). Some studies find that a strong positive relationship holds even over several weeks. The evidence is more mixed on whether *lagged* order flow has predictive value for exchange rates. Because most investors are unlikely to have access to FX dealer flow data the instant that customer trades are initiated, from an investor's perspective the important question is whether lagged order flow data can help in predicting the short-term path that exchange rates will take. The evidence to date is based on very limited data sets from only a few dealers and covering only short periods. Thus, the evidence is too fragmented to draw any firm conclusion on this question.

9.2.2. Extracting Information from the Currency Options Market

FX traders often use currency risk reversals to glean information on whether the FX market might be attaching a higher probability to a large currency appreciation than to a large currency depreciation, or vice versa. A **risk reversal** is a currency option position that consists of the purchase of an out-of-the-money (25 delta) call and the simultaneous sale of an out-of-the-money (25 delta) put, both on the base currency in the P/B exchange rate quote and both with

the same expiration date.[24] Risk reversals are quoted in terms of the implied volatility spread between the 25 delta call and 25 delta put.[25] For example, if the price of the call implies an exchange rate volatility 2 percent larger than the implied volatility built into the put price, the risk reversal would be quoted at +2 percent. If the implied volatility on the put were 2 percent greater than the implied volatility on the call, the risk reversal would be quoted at −2 percent.

A risk reversal quoted at +2 percent (i.e., higher volatility in the call option price) would indicate that the market was attaching a higher probability to a large appreciation of the base currency than to a large depreciation. This would indicate that the market was willing to pay more to insure against the risk that the base currency will rise sharply than it was willing to pay to insure against the risk that the base currency will fall sharply.

The key issue for traders and investors is whether the level or trend in currency risk reversals can be used to correctly anticipate future exchange rate movements. The evidence indicates that there exists a high, contemporaneous correlation between the trend in risk reversals and the trend in exchange rates, but no statistically significant relationship exists between lagged risk reversal data and future exchange rate movements. Therefore, *risk reversals are capable of confirming an exchange rate's trend but cannot predict it*.

9.2.3. Information in the Size and Trend in Net Speculative Positions

FX market participants closely monitor weekly changes in net positions of speculative accounts in the FX futures market to (1) glean whether speculative flows are moving into or out of particular currencies, which would indicate whether speculative flows were exerting significant upward or downward pressure on those currencies or (2) assess whether such positions might be overbought or oversold. If speculative positions were overstretched, it might raise the probability that an unforeseen event or shock could prompt an unwinding of those overstretched positions and, in the process, cause a major reversal in the prevailing exchange rate trend.

The weekly "Commitments of Traders" report of the Commodity Futures Trading Commission (CFTC) contains data on long and short positions held by commercial and noncommercial (speculative) accounts in currency futures contracts that trade on the Chicago Mercantile Exchange. Analysts and market participants often focus on the level and trend in CFTC data on net positions of speculators to assess whether trends in investor positioning offer any insight into the likely direction that exchange rates might take in the future. Recent studies by the Federal Reserve Bank of New York and the Bank of England[26] do indeed find a strong, positive *contemporaneous* relationship between exchange rate movements and changes in net positions of speculative accounts. That is, a buildup of long speculative positions in a particular currency tends to be associated with an appreciation of that currency, and vice versa. These two studies, however, find that *changes in net speculative positions do not lead (i.e., predict) changes in exchange rates*. Nor do extreme overbought or oversold positions correctly

[24]The delta of an option reflects the sensitivity of its price to the price of the underlying instrument. For standard put and call options, delta ranges (in absolute value) between zero (deep out-of-the-money options) and 1 (prices of deep in-the-money options change one-for-one with the underlying price), with at-the-money options having a delta of 0.5. In market jargon, delta is often multiplied by 100, so a delta of 0.25 is referred to as "25 delta."

[25]The values of standard put and call options increase as the volatility of the underlying asset price increases. Given an option pricing model such as the Black–Scholes model, the price of an option can be quoted in terms of the volatility implied by the market price.

[26]Klitgard and Weir (2004) and Mogford and Pain (2006).

anticipate major currency reversals. As we found in the case of FX options, the market simply does not tip its hand ahead of time as to what direction it intends to take.

EXAMPLE 10-14 Technical Analysis and Flow, Sentiment, and Positioning Indicators

Monique Kwan turns her attention to forming short-term currency forecasts based on technical analysis as well as flow, sentiment, and positioning indicators. Her research associate tells her that both net speculative positioning indicators and options market indicators have done a good job of predicting exchange rate trends and market reversals over time.

1. The research associate is *most likely* incorrect:
   A. only with regard to options market indicators.
   B. only with regard to speculative positioning indicators.
   C. with regard to both speculative positioning and options market indicators.

After investigating further, Kwan concludes that technical trading rules can be useful to investors for controlling downside risk, even if such trading rules fail to boost long-run total returns. She bases her conclusion on two findings. First, following the dictates of a technically based trading system may help investors avoid taking extreme contrarian positions. Second, a trend-following trading system overlaid on a carry trade strategy could keep investors from leaning too heavily the wrong way when carry trade unwinds are occurring.

2. Kwan is *most likely* correct:
   A. in both of her findings.
   B. only in regard to her first finding.
   C. only in regard to her second finding.

Kwan examines the order flow for the JPY/USD currency pair passing though the foreign exchange dealer she works for. She finds that net purchases of yen (against the dollar) have significantly exceeded net sales. Based on this finding, she draws two conclusions: first, that there are contemporaneous downward pressures on the JPY/USD spot rate, and second, that the JPY/USD spot rate will trend lower over the next few weeks.

3. Based on historical evidence, Kwan is *most likely* correct:
   A. in both of her conclusions.
   B. only in regard to her first conclusion.
   C. only in regard to her second conclusion.

Solution to 1: C is correct. Studies find that although both speculative positioning indicators and options-market-based indicators may have a *contemporaneous* correlation with exchange rate movements, they have weak predictive power.

Solution to 2: A is correct. If properly deployed, technical trading rules can help manage both the risk of excessive positions as well as the crash risks of carry trades.

Solution to 3: B is correct. The order flow shows net buying of the yen against the dollar, which means *contemporaneous* downward pressure on the JPY/USD currency pair. However, most studies find only weak predictive power from such order flow data.

10. SUMMARY

Exchange rates are one of the most difficult financial market prices to understand and therefore to forecast. There simply is no simple, robust framework—something akin to discounted cash flows in the valuation of equities and fixed-income instruments—on which investors can rely in assessing the appropriate level and likely movements of exchange rates. Nonetheless, ongoing globalization makes it increasingly important for investors to make informed judgments about exchange rates.

In this chapter, we have described the various theories and modeling approaches that economists and foreign exchange strategists have devised to help explain and (hopefully) profit from exchange rate movements. On a theoretical level, we have described how changes in monetary policy, fiscal policy, current account trends, and capital flows affect exchange rate trends, as well as what role central bank intervention and capital controls can play in counteracting potentially undesirable exchange rate movements. The chapter discusses the empirical evidence regarding the ability of our theoretical models to explain and predict exchange rate movements. The reader should have developed an understanding of the fundamental and technical forces that affect exchange rates over short-, medium- and long-run periods, as well as an appreciation of the difficulties one is likely to face in devising a successful and profitable exchange rate forecasting and trading strategy.

This chapter makes the following points, among others:

- Spot exchange rates apply to trades for the next settlement date (usually T + 2) for a given currency pair. Forward exchange rates apply to trades to be settled at any longer maturity.
- Market makers quote bid and offer prices (in terms of the *price currency*) at which they will buy or sell the *base currency.*
  - The offer price is always higher than the bid price.
  - The counterparty that asks for a two-sided price quote has the option (but not the obligation) to deal at either the bid or the offer price quoted.
  - The bid/offer spread depends on (1) the currency pair involved, (2) the time of day, (3) market volatility, (4) the transaction size, and (5) the relationship between the dealer and client. Spreads are tightest in highly liquid currency pairs (e.g., USD/EUR), when the key market centers (e.g., London) are open, and when market volatility is relatively low.
- Absence of arbitrage requires the following:
  - The bid shown by a dealer in the interbank market cannot be higher than the current interbank offer price, and the dealer's offer cannot be lower than the interbank bid.
  - The cross-rate bids posted by a dealer must be lower than the implied cross-rate offers available in the interbank market, and the dealer's offers must be higher than the cross-rate bids. If not, then a triangular arbitrage opportunity arises.
- Forward exchange rates are quoted in terms of points to be added to the spot exchange rate. If the points are positive (or negative), the base currency is trading at a forward premium (or discount). The points are proportional to the interest rate differential and approximately proportional to the time to maturity.
- Forecasting the direction of exchange rate movements can be a daunting task. Most studies find that models that work well in one period or for one set of exchange rates fail to work well for others.
- International parity conditions show us how expected inflation, interest rate differentials, forward exchange rates, and expected future spot exchange rates are linked in an ideal world. According to theory, relative expected inflation rates should determine relative nominal interest rates; relative interest rates, in turn, should determine forward exchange

rates; and forward exchange rates should correctly anticipate the path of the future spot exchange rate.

- International parity conditions tell us that countries with high (or low) expected inflation rates should see their currencies depreciate (or appreciate) over time, that high-yield currencies should see their currencies depreciate relative to low-yield currencies over time, and that forward exchange rates should function as unbiased predictors of future spot exchange rates.
- With the exception of covered interest rate parity, which is enforced by arbitrage, the key international parity conditions rarely hold in either the short or the medium term. However, the parity conditions tend to hold over relatively long horizons.
- According to the theory of covered interest rate parity, an investment in a foreign-currency-denominated money market investment that is completely hedged against exchange rate risk in the forward market should yield exactly the same return as an otherwise identical domestic money market investment.
- According to the theory of uncovered interest rate parity, the expected change in a domestic currency's value should be fully reflected in domestic–foreign interest rate spreads. If the uncovered interest rate parity condition always held, it would rule out the possibility of earning excess returns from going long a high-yield currency and short a low-yield currency.
- According to the *ex ante* purchasing power parity condition, expected changes in exchange rates should equal the difference in expected national inflation rates.
- Most studies find that high-yield currencies do not depreciate and low-yield currencies do not strengthen as much as yield spreads would suggest over short- to medium-term periods. Many investors exploit this anomaly by engaging in so-called carry trades that overweight high-yield currencies at the expense of low-yield currencies. Historically, such carry trades have generated attractive excess returns in benign market conditions but tend to perform poorly when market conditions are highly volatile (i.e., they are subject to crash risk).
- If both *ex ante* purchasing power parity and uncovered interest rate parity held, real interest rates across all markets would be the same. This is real interest rate parity. Combining real interest rate parity with the fact that each country's nominal interest rate equals its real interest rate plus its expected inflation rate, we have the international Fisher effect: the nominal interest rate differential between two currencies equals the difference between the expected inflation rates.
- If both covered and uncovered interest rate parity held, the market would set the forward exchange rate equal to the spot exchange rate that is expected to prevail in the future. That is, the forward exchange rate would serve as an unbiased predictor of the future spot exchange rate.
- The purchasing power parity (PPP) approach to assessing long-run fair value probably has the widest following among international economists.
- The macroeconomic balance approach to assessing long-run fair value in the foreign exchange market estimates how much exchange rates will need to adjust to bring a country's current account balance to a sustainable level.
- The external debt sustainability approach to assessing long-run fair value in the foreign exchange market estimates what exchange rate level will ensure that a country's net external asset or liability position stabilizes at a viable level.
- A useful model of longer-term exchange rate determination can be obtained by combining convergence to a long-run equilibrium real exchange rate with uncovered interest rate parity:

$$q_{f/d} = \overline{q}_{f/d} + [(r_d - r_f) - (\varphi_d - \varphi_f)]$$

- For the most part, countries that run persistent current account deficits will see their currencies weaken over time. Similarly, countries that run persistent current account surpluses will tend to see their currencies appreciate over time.
- The relationship between current account imbalances and changes in exchange rates is not contemporaneous. Indeed, large current account imbalances can persist for long periods of time before they trigger an adjustment in exchange rates.
- A significant adjustment in exchange rates is often required to facilitate correction of a large current account gap. Many studies find long lags, perhaps lasting several years, between (1) the onset of the exchange rate change and (2) the adjustment in traded goods prices in response to the change in the exchange rate, and then (3) the eventual effect of the change in traded goods prices on import and export demand.
- Greater financial integration of the world's capital markets and the increased freedom of capital to flow across national borders have increased the importance of global capital flows in determining exchange rates.
- Countries that run relatively tight monetary policies, introduce structural economic reforms, and lower outsized budget deficits will often see their currencies strengthen over time as capital flows respond positively to relatively high nominal interest rates, lower inflation expectations, a lower risk premium, and an upward revision in the market's assessment of what exchange rate level constitutes long-run fair value.
- Monetary policy affects the exchange rate through a variety of channels. In the Mundell–Fleming model, it does so primarily through the interest rate sensitivity of capital flows, strengthening the currency when monetary policy is tightened and weakening it when monetary policy is eased. The more sensitive capital flows are to the change in interest rates, the greater the exchange rate's responsiveness to the change in monetary policy.
- In the monetary model of exchange rate determination, monetary policy is deemed to have a direct impact on the actual and expected path of inflation, which, via purchasing power parity, translates into a corresponding impact on the exchange rate.
- Although monetary policy impulses may be transmitted to exchange rates through a variety of channels, the end result is broadly the same—countries that pursue overly easy monetary policies will see their currencies depreciate over time. If a central bank wishes to slow or reverse a decline in the value of its currency, a move toward a tighter monetary policy would be helpful, if not required.
- Fiscal policy has an ambiguous impact on the exchange rate. In the Mundell–Fleming model, an expansionary fiscal policy typically results in a rise in domestic interest rates and an increase in economic activity. The rise in domestic interest rates should induce a capital inflow, which is positive for the domestic currency, but the consequent rise in economic activity should contribute to a deterioration of the trade balance, which is negative for the domestic currency. The more mobile capital flows are, the greater the likelihood that the induced inflow of capital will dominate the deterioration in trade.
- Under conditions of high capital mobility, countries that simultaneously pursue expansionary fiscal policies and relatively tight monetary policies should see their currencies strengthen over time.
- The portfolio balance model of exchange rate determination asserts that a steady increase in the stock of government debt outstanding, perhaps generated by a steady widening of the government budget deficit over time, will be willingly held by investors only if they are compensated in the form of a higher expected return The higher expected return could come from (1) higher interest rates and/or higher risk premiums, (2) depreciation of the

currency to a level sufficient to generate anticipation of gains from subsequent currency appreciation, or (3) some combination of the two.

- Surges in capital inflows can be a curse if they fuel boomlike conditions, asset price bubbles, and an overshoot of exchange rates into overvalued territory. One of the major issues confronting policy makers in emerging market countries is how best to respond to excessive surges in capital flows.
- The International Monetary Fund now considers capital controls to be a legitimate part of a policy maker's tool kit. Given the painful lessons from previous episodes of surging capital flows, the IMF feels that under certain circumstances, capital controls may be needed to prevent exchange rates from overshooting, asset price bubbles from forming, and future financial conditions from deteriorating.
- The evidence indicates that intervention by industrial countries has had an insignificant impact on the course of exchange rates. The evidence is more mixed for emerging markets. Emerging market policy makers might have greater success in managing their exchange rates given their large arsenal of foreign exchange reserve holdings, which appear sizable relative to the limited turnover of FX transactions in many emerging markets.
- Although each currency crisis episode is distinct in some respects, an IMF study of 50 episodes found the following stylized facts:
  - Leading up to a crisis, the real exchange rate is substantially higher than its mean level during tranquil periods.
  - The trade balance does not signal an impending currency crisis.
  - Foreign exchange reserves tend to decline precipitously as the crisis approaches.
  - On average, the terms of trade deteriorate somewhat leading up to a crisis.
  - Inflation tends to be significantly higher in precrisis periods.
  - The ratio of M2 (a measure of money supply) to bank reserves tends to rise in the 24-month period leading up to a crisis, then plummets sharply in the months immediately following a crisis.
  - Broad money growth in nominal and real terms tends to rise sharply in the two years leading up to a currency crisis, peaking around 18 months before a crisis hits.
  - Nominal private credit growth tends to rise sharply in the period leading up to a crisis.
  - Currency crises are often preceded by a boom–bust cycle in financial asset (equity) prices.
  - Real economic activity does not display any distinctive pattern ahead of a crisis but falls sharply in the aftermath of a crisis.
- Technical analysis is a popular trading tool for many, if not most, FX market participants. Numerous academic studies conducted in the 1970s, 1980s, and early 1990s concluded that a variety of trend-following trading rules would have generated significant profits had such models been actively followed during that period. However, updated studies for the post-1995 period indicate that trend-following trading rules have not fared as well since.
- Although technical analysis may now be less useful as a strategic tool to enhance return, a number of studies show that technical analysis may be a useful tool in managing the downside risk associated with FX portfolios.
- Most studies find that there exists a strong positive, *contemporaneous* relationship between cumulative order flow and exchange rates over short periods of time. However, the evidence is more mixed regarding whether order flow has *predictive* value for exchange rates.
- Empirical studies find that neither the data on currency risk reversals nor data on the size and trend in reported net speculative positions on the futures market are useful for currency forecasting purposes.

11. APPENDIX: CURRENCY CODES USED IN THIS CHAPTER

USD	U.S. dollar
EUR	Euro
GBP	UK pound
JPY	Japanese yen
MXN	Mexican peso
CHF	Swiss franc
CAD	Canadian dollar
SEK	Swedish krona
AUD	Australian dollar
KRW	South Korean won
NZD	New Zealand dollar

PRACTICE PROBLEMS

The following information relates to Questions 1 through 6.[27]

Ed Smith is a new trainee in the foreign exchange (FX) services department of a major global bank. Smith's focus is to assist the senior FX trader, Feliz Mehmet, CFA. Mehmet mentions that an Indian corporate client exporting to the United Kingdom wants to estimate the potential hedging cost for a sale closing in one year. Smith is to determine the premium or discount for an annual (360-day) forward contract using the exchange rate data presented in Exhibit A:

Exhibit A: Select Currency Data for GBP and INR

Spot (INR/GBP)	79.5093
Annual (360-day) LIBOR (GBP)	5.43%
Annual (360-day) LIBOR (INR)	7.52%

Mehmet is also looking at two possible trades to determine their profit potential. The first trade involves a possible triangular arbitrage trade using the Swiss, U.S., and Brazilian currencies, to be executed based on a dealer's bid/offer rate quote of 0.5161/0.5163 in CHF/BRL and the interbank spot rate quotes presented in Exhibit B:

[27]These practice problems were developed by Greg Gocek, CFA (Downers Grove, Illinois, USA).

Exhibit B: Interbank Market Quotes

Currency Pair	Bid/Offer
CHF/USD	0.9099/0.9101
BRL/USD	1.7790/1.7792

Mehmet is also considering a carry trade involving the USD and the euro. He anticipates it will generate a higher return than buying a one-year domestic note at the current market quote due to low U.S. interest rates and his predictions of exchange rates in one year. To help Mehmet assess the carry trade, Mehmet provides Smith with selected current market data and his one-year forecasts in Exhibit C:

Exhibit C: Spot Rates and Interest Rates for Proposed Carry Trade

Today's One-Year LIBOR		Currency Pair (Price/Base)	Spot Rate Today	Projected Spot Rate in One Year
USD	0.80%	CAD/USD	1.0055	1.0006
CAD	1.71%	EUR/CAD	0.7218	0.7279
EUR	2.20%			

Finally, Mehmet asks Smith to assist with a trade involving a U.S. multinational customer operating in Europe and Japan. The customer is a very cost-conscious industrial company with an AA credit rating and strives to execute its currency trades at the most favorable bid/offer spread. Because its Japanese subsidiary is about to close on a major European acquisition in three business days, the client wants to lock in a trade involving the Japanese yen and the euro as early as possible the next morning, preferably by 8:05 a.m. New York time.

At lunch, Smith and other FX trainees discuss how best to analyze currency market volatility from ongoing financial crises. The group agrees that a theoretical explanation of exchange rate movements, such as the framework of the international parity conditions, should be applicable across all trading environments. They note such analysis should enable traders to anticipate future spot exchange rates. But they disagree on which parity condition best predicts exchange rates, voicing several different assessments. Smith concludes the discussion on parity conditions by stating to the trainees:

> "I believe that in the current environment both covered and uncovered interest rate parity conditions are in effect."

The conversation next shifts to exchange rate assessment tools, specifically the techniques of the IMF Consultative Group on Exchange Rate Issues (CGER). CGER uses a three-part approach that includes the macroeconomic balance approach, the external sustainability approach, and a reduced-form econometric model. Smith asks Leslie Jones, another trainee, to describe the three approaches. In response, Jones makes the following statements to the other trainees and Smith:

Statement 1: "The macroeconomic balance approach focuses on the stocks of outstanding assets and liabilities."

Statement 2: "The reduced-form econometric model has a weakness in underestimating future appreciation of undervalued currencies."

Statement 3: "The external sustainability approach centers on adjustments leading to long-term equilibrium in the capital account."

1. Based on Exhibit A, the forward premium or discount for a 360-day INR/GBP forward contract is *closest* to:
   A. −1.546.
   B. 1.546.
   C. 1.576.

2. Based on Exhibit B, the *most* appropriate recommendation regarding the triangular arbitrage trade is to:
   A. decline the trade, as no arbitrage profits are possible.
   B. execute the trade; buy BRL in the interbank market and sell it to the dealer.
   C. execute the trade; buy BRL from the dealer and sell it in the interbank market.

3. Based on Exhibit C, the potential all-in USD return on the carry trade is *closest* to:
   A. 1.04 percent.
   B. 1.40 percent.
   C. 1.84 percent.

4. The factor *least likely* to lead to a narrow bid/offer spread for the industrial company's needed currency trade is:
   A. the timing of its trade.
   B. the company's credit rating.
   C. the pair of currencies involved.

5. If Smith's statement on parity conditions is correct, future spot exchange rates are *most likely* to be forecast by:
   A. current spot rates.
   B. forward exchange rates.
   C. inflation rate differentials.

6. Which of the following statements given by trainee Jones in describing the approaches used by CGER is *most* accurate?
   A. Statement 1
   B. Statement 2
   C. Statement 3

The following information relates to Questions 7 through 13.[28]

Connor Wagener, a student at the University of Canterbury in New Zealand, has been asked to prepare a presentation on foreign exchange rates for his international business course.

[28]These practice problems were developed by Sue Ryan, CFA (East Hartford, Connecticut, USA).

Wagener has a basic understanding of exchange rates, but would like a practitioner's perspective, and he has arranged an interview with currency trader Hannah McFadden. During the interview, Wagener asks McFadden:

> "Could you explain what drives exchange rates? I'm curious as to why our New Zealand dollar was affected by the European debt crisis in 2011 and what other factors impact it."

In response, McFadden begins with a general discussion of exchange rates. She notes that international parity conditions illustrate how exchange rates are linked to expected inflation, interest rate differences, and forward exchange rates as well as current and expected future spot rates. McFadden states:

> ***Statement 1:*** "Fortunately, the international parity condition most relevant for FX carry trades does not always hold."

McFadden continues her discussion:

> "FX carry traders go long (i.e., buy) high-yield currencies and fund their positions by shorting—that is, borrowing in—low-yield currencies. Unfortunately, crashes in currency values can occur, which create financial crises as traders unwind their positions. For example, in 2008, the New Zealand dollar was negatively impacted when highly leveraged carry trades were unwound. In addition to investors, consumers and business owners can also affect currency exchange rates through their impact on their country's balance of payments. For example, if New Zealand consumers purchase more goods from China than New Zealand businesses sell to China, New Zealand will run a trade account deficit with China."

McFadden further explains:

> ***Statement 2:*** "A trade surplus will tend to cause the currency of the country in surplus to appreciate, whereas a deficit will cause currency depreciation. Exchange rate changes will result in immediate adjustments in the prices of traded goods as well as in the demand for imports and exports. These changes will immediately correct the trade imbalance."

McFadden next addresses the influence of monetary and fiscal policy on exchange rates:

> "Countries also exert significant influence on exchange rates both through the initial mix of their fiscal and monetary policies and also by subsequent adjustments to those policies. Various models have been developed to identify how these policies affect exchange rates. The Mundell-Fleming model addresses how changes in both fiscal and monetary policies affect interest rates and ultimately exchange rates in the short term."

McFadden describes monetary models by stating:

> ***Statement 3:*** "Monetary models of exchange rate determination focus on the effects of inflation, price level changes, and risk premium adjustments."

McFadden continues her discussion:

"So far, we've touched on balance of payments and monetary policy. The portfolio balance model addresses the impacts of sustained fiscal policy on exchange rates. I must take a client call, but will return shortly. In the meantime, here is some relevant literature on the models I mentioned along with a couple of questions for you to consider."

Question 1: Assume an emerging market (EM) country has restrictive monetary and fiscal policies under low capital mobility conditions. Are these policies likely to lead to currency appreciation or currency depreciation, or to have no impact?

Question 2: Assume a developed market (DM) country has an expansive fiscal policy under high capital mobility conditions. Why is its currency most likely to depreciate in the long run under an integrated Mundell-Fleming and portfolio balance approach?

Upon her return, Wagener and McFadden review the questions. McFadden notes that capital flows can have a significant impact on exchange rates and have contributed to currency crises in both EM and DM countries. She explains that central banks, like the Reserve Bank of New Zealand, use FX market intervention as a tool to manage exchange rates. McFadden states:

Statement 4: "Some studies have found that EM central banks tend to be more effective in using exchange rate invention than DM central banks, primarily because of one important factor."

McFadden continues her discussion:

Statement 5: "I mentioned that capital inflows could cause a currency crisis, leaving fund managers with significant losses. In the period leading up to a currency crisis, I would predict that an affected country's:
Prediction 1: foreign exchange reserves will increase.
Prediction 2: broad money growth in nominal and real terms will increase.
Prediction 3: real exchange rate will be substantially higher than its mean level during tranquil periods."

After the interview, McFadden agrees to meet the following week with Wagener to discuss more recent events affecting the New Zealand dollar.

7. The international parity condition McFadden is referring to in Statement 1 is:
   A. purchasing power parity.
   B. covered interest rate parity.
   C. uncovered interest rate parity.

8. In Statement 2, McFadden is *most likely* failing to consider:
   A. the initial gap between the country's imports and exports.
   B. the price elasticity of export demand versus import demand.
   C. the lag in the response of import and export demand to price changes.

9. The *least* appropriate factor used to describe the type of models mentioned in Statement 3 is:
   A. inflation.
   B. price level changes.
   C. risk premium adjustments.

10. The best response to Question 1 is that the policies will:
   A. have no impact.
   B. lead to currency appreciation.
   C. lead to currency depreciation.

11. The most likely response to Question 2 is a(n):
   A. increase in the price level.
   B. decrease in risk premiums.
   C. increase in government debt.

12. The factor that McFadden is *most likely* referring to in Statement 4 is:
   A. FX reserve levels.
   B. domestic demand.
   C. the level of capital flows.

13. Which of McFadden's predictions in Statement 5 is *least* correct?
   A. Prediction 1
   B. Prediction 2
   C. Prediction 3

CHAPTER 11

ECONOMIC GROWTH AND THE INVESTMENT DECISION

Paul Kutasovic, CFA

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Describe and compare factors favoring and limiting economic growth in developed and developing economies.
- Describe the relationship between the long-run rate of stock market appreciation and the sustainable growth rate of the economy.
- Explain the importance of potential gross domestic product (GDP) and its growth rate in the investment decisions of equity and fixed-income investors.
- Distinguish between capital deepening investment and technological process and explain the impact of each on economic growth and labor productivity.
- Forecast potential GDP based on growth accounting relations.
- Explain the impact of natural resources on economic growth, and evaluate the argument that limited availability of natural resources constrains economic growth.
- Explain the effects of demographics, immigration, and labor force participation on the rate and sustainability of economic growth.
- Explain how investment in physical capital, human capital, and technological development affects economic growth.
- Compare classical growth theory, neoclassical growth theory, and endogenous growth theory.
- Explain and evaluate convergence hypotheses.
- Explain the economic rationale for governments to provide incentives to private investment in technology and knowledge.
- Describe the expected impact of removing trade barriers on capital investment and profits, employment and wages, and growth in the economies involved.

1. INTRODUCTION

Forecasts of long-run economic growth are important for global investors. Equity prices reflect expectations of the future stream of earnings, which depend on expectations of future economic activity. This means that in the long term, the same factors that drive economic growth will be reflected in equity values. Similarly, the expected long-run growth rate of real income is a key determinant of the average real interest rate level in the economy, and therefore the level of real returns in general. In the shorter term, the relationship between actual and potential growth (i.e., the degree of slack in the economy) is a key driver of fixed-income returns. Therefore, in order to develop global portfolio strategies and investment return expectations, investors must be able to identify and forecast the factors that drive long-term sustainable growth trends. Based on a country's long-term economic outlook, investors can then evaluate the long-term investment potential and risk of investing in the securities of companies located or operating in that country.

In contrast to the short-run fluctuations of the business cycle, the study of economic growth focuses on the long-run trend in aggregate output as measured by potential GDP. Over long periods of time, the actual growth rate of GDP should equal the rate of increase in potential GDP because, by definition, output in excess of potential GDP requires employing labor and capital beyond their optimum levels. Thus, the growth rate of potential GDP acts as an upper limit to growth and determines the economy's sustainable rate of growth. Increasing the growth rate of potential GDP is the key to raising the level of income, the level of profits, and the living standard of the population. Even small differences in the growth rate translate into large differences in the level of income over time.

What drives long-run growth? What distinguishes the winners from the losers in the long-run growth arena? Will poor countries catch up with rich countries over time? Can policies have a permanent effect on the sustainable growth rate? If so, how? If not, why not? These and other key questions are addressed in detail in this chapter.

The chapter is organized as follows: Section 2 examines the long-term growth record, focusing on the extent of growth variation across countries and across decades. Section 3 discusses the importance of economic growth to global investors and examines the relationship between investment returns and economic growth. Section 4 examines the factors that determine long-run economic growth. Section 5 presents the classical, neoclassical, and endogenous growth models. It also discusses whether poorer countries are converging to the higher income levels of the richer countries. Finally, Section 6 looks at the impact of international trade on economic growth. A summary and practice problems complete the chapter.

2. GROWTH IN THE GLOBAL ECONOMY: DEVELOPED VERSUS DEVELOPING COUNTRIES

The first step in our study of long-term growth is to compare the economic performance of countries. GDP and per capita GDP are the best indicators economists have for measuring a country's standard of living and its level of economic development. Economic growth is calculated as the annual percentage change in real GDP or in real per capita GDP. Growth in real GDP measures how rapidly the total economy is expanding. Real per capita GDP reflects the average standard of living in each country—essentially the average level of material well-being. Growth in real GDP per capita (i.e., real GDP growing faster than the population) implies a rising standard of living.

Exhibit 11-1 presents data on the level of per capita GDP and the growth rate of GDP for various countries. Because each country reports its data in its own currency, each country's data must be converted into a common currency, usually the U.S. dollar. One can convert the GDP data into dollars using either current market exchange rates or the exchange rates implied by **purchasing power parity (PPP)**. Purchasing power parity is the idea that exchange rates move to equalize the purchasing power of different currencies. At the exchange rates implied by PPP, the cost of a typical basket of goods and services is the same across all countries. In other words, exchange rates should be at a level where you can buy the same goods and services with the equivalent amount of any country's currency.

EXHIBIT 11-1 Divergent Real GDP Growth among Countries

	Average Annual Real GDP Growth (percent)				Real GDP Per Capita in Dollars[a]			
	1971–1980	1981–1990	1991–2000	2001–2010	1950	1970	1990	2010
Advanced Economies	3.2	3.1	2.8	1.6				
Canada	4.0	2.8	2.4	1.8	$12,053	$19,919	$31,969	$41,288
United States	3.1	2.9	3.4	1.6	14,559	22,806	35,328	46,697
France	2.9	2.4	1.7	1.2	8,266	18,186	28,127	34,358
Germany	2.7	2.3	2.3	0.9	NA	NA	28,624	37,367
Ireland	4.7	3.9	7.1	2.6	5,496	9,869	18,812	36,433
Italy	3.4	2.2	1.7	0.3	5,954	16,522	27,734	31,069
Spain	3.0	3.0	2.9	2.1	3,964	11,444	21,830	30,504
United Kingdom	1.6	2.7	2.8	1.5	11,602	18,002	27,469	37,378
Hong Kong	9.2	6.5	3.9	4.1	3,128	8,031	24,734	43,324
Japan	4.3	4.0	1.3	0.8	3,048	15,413	29,813	34,828
Singapore	10.5	7.3	7.3	5.6	4,299	8,600	27,550	56,224
South Korea	7.4	9.1	7.2	4.2	1,185	3,009	12,083	30,079
Taiwan	10.9	7.9	6.5	3.8	1,425	3,948	15,465	36,413
Australia	3.2	3.3	3.4	3.0	13,219	21,444	30,628	45,951
New Zealand	1.6	2.5	2.9	2.3	13,795	18,255	22,331	31,223
Developing Countries	4.3	4.2	5.4	6.3				
Developing Asia	6.2	6.9	7.4	8.5				
China	10.4	9.1	10.4	10.5	402	698	1,677	8,569
India	3.9	5.9	5.6	7.5	658	922	1,390	3,575
Indonesia	8.4	5.4	4.0	5.2	804	1,182	2,517	4,740

(*Continued*)

EXHIBIT 11-1 *Continued*

	Average Annual Real GDP Growth (percent)				Real GDP Per Capita in Dollars[a]			
	1971–1980	1981–1990	1991–2000	2001–2010	1950	1970	1990	2010
Pakistan	4.5	6.0	3.9	4.8	666	985	1,645	2,600
Philippines	6.6	1.7	3.0	4.7	1,296	2,136	2,660	3,672
Vietnam	4.7	5.9	7.6	7.3	689	770	1,073	3,369
Middle East	2.9	3.0	4.0	4.9				
Egypt	5.9	5.9	4.4	4.9	1,132	1,560	3,137	5,306
Turkey	4.1	5.2	3.6	4.0	2,327	4,413	7,741	11,769
Saudi Arabia	11.0	1.7	2.7	3.3	5,060	17,292	20,399	22,951
Latin America	6.5	1.6	3.3	3.4				
Argentina	2.9	−1.2	4.2	4.6	6,164	9,026	7,952	13,468
Brazil	8.8	1.5	2.5	3.6	2,365	4,324	6,959	9,589
Mexico	6.6	1.8	3.5	1.8	4,180	7,634	10,754	13,710
Peru	7.6	−0.8	4.0	5.7	3,464	5,786	4,516	8,671
Venezuela	1.6	1.9	2.1	3.5	8,104	11,590	9,028	10,560
Africa	3.5	2.5	2.4	5.7				
Botswana	17.1	10.9	6.4	4.2	449	774	3,731	5,311
Ethiopia	3.0	1.9	2.9	8.4	314	479	462	749
Kenya	7.4	4.3	1.7	4.1	791	1,113	1,359	1,376
Nigeria	7.4	2.0	1.9	8.7	814	1,183	1,203	2,037
South Africa	4.1	1.5	1.8	3.5	4,361	6,959	6,595	8,716

[a]The measure of GDP per capita is in constant U.S. dollar market prices for 2010 and adjusted for cross-country differences in the relative prices of goods and services using purchasing power parity (PPP).

Sources: International Monetary Fund, *World Economic Outlook* database for growth rates, and Conference Board Total Economy Database (September 2011).

In general, the simple method of taking a country's GDP measured in its own currency and then multiplying by the current exchange rate to express it in another currency is not appropriate. Using market exchange rates has two problems. First, market exchange rates are very volatile. Changes in the exchange rate could result in large swings in measured GDP even if there is little or no growth in the country's economy. Second, market exchange rates are determined by financial flows and flows in tradable goods and services. This ignores the fact that much of global consumption is for nontradable goods and services. Prices of nontraded goods and services differ by country. In particular, nontraded goods are generally less

expensive in developing countries than in developed countries. For example, because labor is cheaper in Mexico City than in London, the prices of labor-intensive products, such as haircuts or taxi rides, are lower in Mexico City than in London. Failing to account for differences in the prices of nontraded goods and services across countries tends to understate the standard of living of consumers in developing countries. To compare standards of living across time or across countries, we need to use a common set of prices for a wide range of goods and services. Thus, cross-country comparisons of GDP should be based on purchasing power parity rather than current market exchange rates.

The countries in Exhibit 11-1 are divided into two categories, advanced (developed) economies and developing countries. Developed countries are those with high per capita GDP.[1] These include the United States, Canada, Australia, Japan, and major economies in Europe. Growth in the large, developed economies generally slowed over the past few decades, with U.S. growth exceeding that of Europe and Japan. Also included in this group are such markets as Taiwan, South Korea, Singapore, Ireland, and Spain, which were poor in the 1950s but now have relatively high per capita real GDPs because of high rates of growth over the past 50 years.

The second group of countries is the developing countries of Africa, Asia, and Latin America. Per capita GDP in these countries is lower than in the advanced countries, but GDP is generally growing at a faster rate than in the developed countries. Although the growth rates of the developing countries exceed those of the advanced countries, there is significant variation in economic performance among the developing countries. China and India are growing at a rapid rate. Between 1991 and 2010, the Chinese and Indian economies expanded at annual rates of 10.5 percent and 6.6 percent, respectively, compared with U.S. growth of 2.5 percent per year over this period. Meanwhile, growth in Latin America, Africa, and the Middle East has lagged behind Asia.

What explains the diverse experiences among the developing countries and between the developed and developing ones? Singapore, for example, had less than half the per capita GDP of the United States in 1970 but now has per capita GDP that exceeds that of the United States. South Korea and Taiwan have gone from among the poorest economies in the world to among the richest in one generation. In contrast, such countries as Ethiopia and Kenya have remained poor, with little growth in per capita GDP. The literature on economic growth focuses primarily on the role of capital and labor resources and the use of technology as sources of growth. In addition to these purely economic drivers, developed and developing countries differ with respect to the presence or absence of appropriate institutions that support growth. These institutions enable developing countries to raise their standards of living and eventually move into the ranks of the developed countries. We now examine some of the key institutions and requirements for growth.

2.1. Savings and Investment

One of the major problems for some of the developing countries is a low level of capital per worker. Countries accumulate capital through private-sector and public-sector (e.g., infrastructure) investment. But increasing the investment rate may be difficult in developing countries because low levels of disposable income can make it difficult to generate significant

[1]There are no universally agreed-upon criteria for classifying countries as advanced or developing. The International Monetary Fund (IMF) classifies 34 countries as advanced and 150 as developing. It says that "this classification is not based on strict criteria, economic or otherwise, and has evolved over time" (IMF 2011b).

saving. The low saving rate contributes to a vicious cycle of poverty: Low savings lead to low levels of investment, which leads to slow GDP growth, which implies persistently low income and savings. Therefore, it is very difficult to design policies to increase domestic saving and investment rates in developing countries. The good news is that the savings of domestic residents are not the only source of investment funds. A developing country can break out of the cycle of low savings by attracting foreign investment.

2.2. Financial Markets and Intermediaries

In addition to the saving rate, growth depends on how efficiently saving is allocated within the economy. A role of the financial sector in any economy is to channel funds from savers to investment projects. Financial markets and intermediaries, such as banks, can promote growth in at least three ways. First, by screening those who seek funding and monitoring those who obtain funding, the financial sector channels financial capital (savings) to projects that are likely to generate the highest risk-adjusted returns. Second, the financial sector may encourage savings and assumption of risk by creating attractive investment instruments that facilitate risk transfer and diversification and enhance liquidity. Finally, the existence of well-developed financial markets and intermediaries can mitigate the credit constraints that companies might otherwise face in financing capital investments. For example, banks can aggregate small amounts of savings into a larger pool enabling them to finance larger projects that can exploit economies of scale. Evidence suggests that countries with better-functioning financial markets and intermediaries grow at a faster rate.[2] However, not all financial sector developments promote economic growth. Financial sector intermediation that results in declining credit standards or increasing leverage will increase risk and not necessarily increase long-run growth.

2.3. Political Stability, Rule of Law, and Property Rights

Stable and effective government, a well-developed legal and regulatory system, and respect for property rights are key ingredients for economic growth. Property rights are the legal arrangements that govern the protection of private property, including intellectual property. Clearly established property rights create the incentive for domestic households and companies to invest and save. A legal system—substantive and procedural laws[3]—is needed to establish and protect these rights. In developed countries these rights and arrangements are well established, but they may be lacking or ineffective in developing countries.

In addition, economic uncertainty increases when wars, military coups, corruption, and other sources of political instability are widespread. These factors raise investment risk, discourage foreign investment, and weaken growth. In many developing countries, especially those in Africa, the first priority in trying to enhance growth is to enact a legal system that establishes, protects, and enforces property rights.

2.4. Education and Health Care Systems

Inadequate education at all levels is a major impediment to growth for many developing countries. Many workers are illiterate, and few workers have the skills needed to use the latest

[2]Levine (2005).

[3]Substantive law focuses on the rights and responsibilities of entities and relationships among entities, and procedural law focuses on the protection and enforcement of the substantive laws.

technology. At the same time, many developing countries also suffer from a so-called brain drain, where the most highly educated individuals leave the developing country for the advanced countries. Basic education raises the skill level of the workforce and thus contributes to the country's potential for growth. In addition, because physical capital and human capital are often complementary, education can raise growth by increasing the productivity of existing physical capital. Thus, improving education, through both formal schooling and on-the-job training, is an important component of a sustainable growth strategy for a developing country. China and India are investing large amounts in education and have successfully graduated large numbers of students majoring in engineering and technology-related areas of study. This effort is significantly improving the quality of their workforces.

Empirical studies show that the allocation of education spending among different types and levels (primary, secondary, and postsecondary) of education is a key determinant of growth, especially in comparing growth in the developed countries with growth in the developing ones. The impact of education spending depends on whether the country is on the leading edge of technology and fostering innovation or simply relying on imitation as a source of growth. Typically, developed countries, such as the United States, Japan, and western European nations, are on the leading edge of technology and need to invest in postsecondary education to encourage innovation and growth. For these countries, incremental spending on primary and secondary education will have a smaller impact on growth. In contrast, the developing countries, which largely apply technology developed elsewhere, should emphasize primary and secondary education. Such spending will improve growth by improving the countries' ability to absorb new technologies and to organize existing tasks more efficiently.

Poor health is another obstacle to growth in the developing countries. Life expectancy rates are substantially lower in many developing countries. In Africa, tropical diseases are rampant and AIDS has had a devastating impact. As is evident in Exhibit 11-1, the growth rate of GDP in Botswana, a huge success story in the 1970s and 1980s, has slowed dramatically over the past two decades due, at least in part, to the AIDS epidemic.

2.5. Tax and Regulatory Systems

Tax and regulatory policies have an important impact on growth and productivity, especially at the company level. Analysis suggests that limited regulations encourage entrepreneurial activity and the entry of new companies. There is also a strong positive correlation between the entry of new companies and average productivity levels. Studies by the Organization for Economic Cooperation and Development (OECD) indicate that low administrative start-up cost is a key factor encouraging entrepreneurship.[4]

2.6. Free Trade and Unrestricted Capital Flows

Opening an economy to capital and trade flows has a major impact on economic growth. In an open economy, world savings can finance domestic investment. As a potential source of funds, foreign investment can break the vicious cycle of low income, low domestic savings, and low investment. Foreign investment can occur in two ways:

1. Foreign companies can invest directly in a domestic economy (so-called foreign direct investment [FDI]) by building or buying property, plant, and equipment.

[4]OECD (2003).

2. Foreign companies and individuals can invest indirectly in a domestic economy by purchasing securities (equity and fixed income) issued by domestic companies.

Both of these forms of foreign investment will potentially increase the developing economy's physical capital stock, leading to higher productivity, employment, and wages, and perhaps even increased domestic savings. This suggests that developing countries would benefit from policies that encourage investment from abroad, such as eliminating high tariffs on foreign imports (especially capital goods) and removing restrictions on foreign direct and indirect investments.

Brazil and India are examples of developing countries that have benefited from foreign investment. Foreign companies directly invested $48.5 billion in Brazil in 2010, an important source of investment spending for the Brazilian economy (see Exhibit 11-19 in Section 6). Foreign direct investment also provides developing countries with access to advanced technology developed and used in the advanced countries. In 1999, India enacted new regulations that liberalized direct and indirect foreign investments in Indian companies. Foreign institutional and venture capital investors were given greater flexibility to invest directly in Indian entities as well as in the Indian capital markets. These changes also made it easier for foreign companies to invest in plant and equipment. These developments contributed to the acceleration in India's economic growth over the past decade (see Exhibit 11-1).

Capital flows are just one way that the international economy affects economic growth. The other is through trade in goods and services. In general, free trade benefits an economy by providing its residents with more goods at lower costs. Domestic companies face increased competition, which limits their price discretion, but they also obtain access to larger markets. The evidence of the benefits of open markets is discussed later in the chapter.

2.7. Summary of Factors Limiting Growth in Developing Countries

Developing countries differ significantly from developed countries in terms of their institutional structures and their legal and political environments. Lack of appropriate institutions and poor legal and political environments restrain growth in the developing economies and partially explain why these countries are poor and experience slow growth. Factors limiting growth include:

- Low rates of saving and investment.
- Poorly developed financial markets.
- Weak, or even corrupt, legal systems and failure to enforce laws.
- Lack of property rights.
- Political instability.
- Poor public education and health services.
- Tax and regulatory polices discouraging entrepreneurship.
- Restrictions on international trade and flows of capital.

Although these factors are not necessarily absent in developed countries, they tend to be more prevalent in developing countries. Policies that correct these issues or mitigate their impact enhance the potential for growth. In addition to these institutional restraints, as we will see in Section 4, growth in developing countries may be limited by a lack of physical, human, and public capital, as well as little or no innovation.

EXAMPLE 11-1 Why Growth Rates Matter*

In 1950, Argentina and Venezuela were relatively wealthy countries with per capita levels of GDP of $6,164 and $8,104, respectively. Per capita GDPs in these Latin American countries were well above those of Japan, South Korea, and Singapore, which had per capita GDPs of $3,048, $1,185, and $4,299, respectively. By 2010, however, a dramatic change had occurred in the relative GDPs per capita of these countries.

Real GDP Per Capita in Dollars

	Venezuela	Argentina	Singapore	Japan	South Korea
1950	$8,104	$6,164	$4,299	$3,048	$1,185
2010	$10,560	$13,468	$56,224	$34,828	$30,079

1. Calculate the annual growth rate in per capita GDP for each of the five countries over the period from 1950 to 2010.
2. Explain the implication of the growth rates for these countries.
3. Suppose that GDP per capita in Argentina had grown at the same rate as in Japan from 1950 to 2010. How much larger would real per capita GDP have been in Argentina in 2010?
4. Venezuela plans to stimulate growth in its economy by substantially increasing spending on infrastructure, education, and health care. Nevertheless, foreign investment is discouraged, and reforms like strengthening the legal system and encouraging private ownership have been largely ignored. Explain whether the measures described earlier could lead to faster economic growth.

Solution to 1: The annual growth rates for the five countries are calculated as follows:

Argentina	$(\$13{,}468/\$6{,}164)^{1/60} - 1 = 1.31\%$
Venezuela	$(\$10{,}560/\$8{,}104)^{1/60} - 1 = 0.44\%$
Japan	$(\$34{,}828/\$3{,}048)^{1/60} - 1 = 4.14\%$
Singapore	$(\$56{,}224/\$4{,}299)^{1/60} - 1 = 4.38\%$
South Korea	$(\$30{,}079/\$1{,}185)^{1/60} - 1 = 5.54\%$

Solution to 2: Differences in GDP growth rates sustained over a number of decades will significantly alter the relative incomes of countries. Nations that experience sustained periods of high growth will eventually become high-income countries and move up the income ladder. In contrast, countries with slow growth will experience relative declines in living standards. This is well illustrated in this example by a historic comparison of growth in Argentina and Venezuela with Japan, Singapore, and South Korea. In 1950, Argentina and Venezuela were relatively wealthy countries with per capita levels of GDP well above

those of Japan, South Korea, and Singapore. Over the next 60 years, however, the rate of growth in per capita GDP was significantly slower in Venezuela and Argentina in comparison to the three Asian countries. This resulted in a dramatic change in the relative incomes of these countries. The per capita GDP of the three Asian countries rose sharply as each joined the ranks of developed countries. In contrast, Argentina and Venezuela stagnated and moved from the ranks of developed countries to developing country status. By 2010, per capita income in Singapore was more than five times higher than in Venezuela.

Over the long run, the rate of economic growth is an extremely important variable. Even small differences in growth rates matter because of the power of compounding. Thus, policy actions that affect the long-term growth rate even by a small amount will have a major economic impact.

Solution to 3: Assuming Argentina had grown at the same rate as Japan since 1950, its GDP per capita in 2010 would have been $(\$6{,}164)(1 + 0.0414)^{60} = (\$6{,}164)(11.404) = \$70{,}294$, versus $13,468 from Exhibit 11-1.

If Argentina had grown at the same rate as Japan, it would have had by far the highest standard of living in the world in 2010. The question is why the growth rates in Argentina and Venezuela diverged so much from the rates of the three Asian countries.

Solution to 4: The preconditions for economic growth are well-functioning financial markets, clearly defined property rights and rule of law, open international trade and flows of capital, an educated and healthy population, and tax and regulatory policies that encourage entrepreneurship. Investment in infrastructure would increase Venezuela's stock of physical capital, which would raise labor productivity and growth. Better education and health care would increase human capital and also increase productivity and growth. These measures would raise the growth prospects for Venezuela. However, what is missing is a legal system that could better enforce property rights, openness to international trade and foreign investment, and well-functioning capital markets. Without changes in these preconditions, a significant improvement in growth is unlikely to occur. The preconditions are summarized here:

Preconditions for Growth	Impact of Planned Policy Action in Venezuela
Saving and investment	Improve growth potential
Developed financial markets	No impact
Legal systems	No impact
Property rights and political stability	No impact
Education and health	Improve growth potential
Tax and regulatory polices discouraging entrepreneurship	No impact
Open international trade and flows of capital	No impact

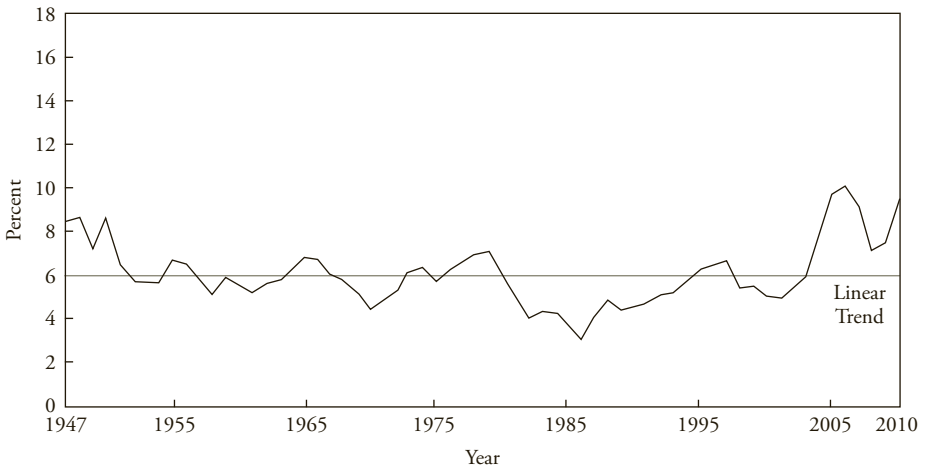
*It should be noted that the global economy is evolving rapidly and past trends may or may not be sustained. Nonetheless, in order to provide concrete answers that do not require the reader to bring in additional information, our exercise solutions must assume past patterns are indicative of the future.

3. WHY POTENTIAL GROWTH MATTERS TO INVESTORS

The valuations of both equity and fixed-income securities are closely related to the growth rate of economic activity. Anticipated growth in aggregate earnings is a fundamental driver of the equity market. Growth in an economy's productive capacity, measured by **potential GDP**, places a limit on how fast the economy can grow. The idea is that potential GDP is the maximum amount of output an economy can sustainably produce without inducing an increase in the inflation rate. A key question for equity investors, therefore, is whether earnings growth is also bounded or limited by the growth rate of potential GDP.

For earnings growth to exceed GDP growth, the ratio of corporate profits to GDP must trend upward over time. It should be clear that the share of profits in GDP cannot rise forever. At some point, stagnant labor income would make workers unwilling to work and would also undermine demand, making further profit growth unsustainable. Thus, in the long run, real earnings growth cannot exceed the growth rate of potential GDP.[5] Exhibit 11-2 illustrates the long-run stability of after-tax profits as a share of GDP using U.S. data derived from the National Income and Product Accounts (NIPA). The chart shows that since 1947, after-tax profits have ranged between 3.1 percent and 10.1 percent of GDP and have averaged around 6 percent of GDP. Note that there is neither an upward trend in the ratio of after-tax profits to GDP nor a move to a permanent increase in the ratio. The share of profits in 1947, at 8.5 percent, was essentially equal to the 9.4 percent share at the end of the period in 2010.

EXHIBIT 11-2 U.S. After-Tax Corporate Profits as a Percentage of GDP



Source: FRED database, Federal Reserve Bank of St. Louis.

[5]Earnings growth for the overall national economy can differ from the growth of earnings per share in a country's equity market composites. This is due to the presence of new businesses that are not yet included in the equity indexes and are typically growing at a faster rate than the mature companies that make up the composites. Thus, the earnings growth rate of companies making up the composites should be lower than the earnings growth rate for the overall economy.

Because there is no trend in the ratio, the same factors that limit economic growth also set the upper limit or bound on the long-run growth of aggregate earnings.

To examine the relationship between economic growth and stock prices, it is useful to express the aggregate value of the stock market as the product of key ratios. Letting P represent the aggregate value (price) of equities and E represent aggregate earnings, we can write:

$$P = GDP\left(\frac{E}{GDP}\right)\left(\frac{P}{E}\right)$$

This equation represents the aggregate value of equities as the product of GDP, corporate earnings as a share of GDP, and the price-to-earnings (P/E) ratio for the market. Note that GDP may be interpreted as either real or nominal, with a corresponding real or nominal interpretation of the other variables.

This equation can be expressed in terms of logarithmic rates of change over a time horizon T:

$$(1/T)\%\Delta P = (1/T)\%\Delta GDP + (1/T)\%\Delta(E/GDP) + (1/T)\%\Delta(P/E)$$

Thus, the percentage change in stock market value equals the percentage change in GDP plus the percentage change in the share of earnings (profit) in GDP plus the percentage change in the price-to-earnings multiple.[6] Over short to immediate horizons, all three of these factors contribute to appreciation or depreciation of the stock market. In the long run, however, the growth rate of GDP must dominate. As noted earlier, the ratio of earnings to GDP cannot rise forever. It cannot decline forever, either, because unprofitable businesses will disappear. Hence, the second term in the equation must be approximately zero over long horizons (T). Similarly, the price-to-earnings ratio cannot grow or contract forever, because investors will not pay an arbitrarily large price for a unit of earnings, nor will they give away earnings for nothing. Hence, the third term must also be approximately zero over long horizons. The conclusion is that the drivers of potential GDP are ultimately the drivers of stock market price performance.

Exhibit 11-3 shows the close relationship between economic growth and equity market appreciation over long horizons. Over the period 1946–2007, the S&P 500 index returned 10.82 percent per year, of which 7.15 percent per year came from price appreciation. The price appreciation was almost exactly equal to the 6.95 percent growth rate of U.S. nominal GDP (real GDP growth plus inflation). Changes in the earnings-to-GDP and price-to-earnings ratios contributed only a combined 0.20 percent per year. As shown in the last column of the exhibit, these two ratios contributed much more to the volatility of the market than to its return.[7]

Estimates of potential GDP and its growth rate are widely available. For example, both the OECD and the International Monetary Fund (IMF) provide such estimates as a basis for

[6]For simplicity, we have not explicitly incorporated issuance or repurchasing of shares. To do so, we would simply need to distinguish between aggregate stock market value and price per share. However, this would not alter our conclusions. Similarly, we could incorporate the dividend payout ratio into our argument, but again, this would not alter our conclusions.

[7]It should be noted that the 1946–2007 time period was chosen because both end points correspond to fairly normal economic and market conditions. Selecting end points that correspond to crisis or bubble conditions would distort the role played by the various components of return.

EXHIBIT 11-3 Decomposition of S&P 500 Returns: Log Returns, 1946–2007

	Annual Return/Growth Rate	Standard Deviation
S&P 500 return	10.82%	15.31%
Real GDP growth	3.01	2.97
Inflation	3.94	3.29
EPS/GDP	−0.12	17.62
P/E	0.32	23.80
Dividend yield	3.67	1.49
Total	10.82	

Source: Stewart, Piros, and Heisler (2011).

their intermediate-term and long-term forecasts of economic growth by country. In addition, central banks regularly make projections of potential GDP. The methods used to estimate potential GDP are examined later in the chapter. The data in Exhibit 11-1 illustrate that simply extrapolating past GDP growth into the future may produce an incorrect forecast. A country's GDP growth rate can and does change over time. GDP growth can either slow down, as was the case for Japan (compare 1971–1990 with 1991–2010), or speed up, as was the case for Brazil over the past decade. Factors or policies that cause potential growth to increase or decrease by even a small amount will have a large impact on living standards and the future level of economic activity. The effect is analogous to the rate of return on a portfolio, where small differences in return compounded over many years result in a substantially higher or lower value for the portfolio. Being able to recognize these changes is critical for the global investor.

Estimates of an economy's growth potential are also relevant for global fixed-income investors. One of the uses of potential GDP is to gauge inflationary pressures in the economy. Actual GDP growth above (or below) the potential growth rate puts upward (or downward) pressure on inflation, which puts corresponding pressure on nominal interest rates and bond prices.[8]

The growth rate of potential GDP is also an important determinant of the level of real interest rates, and therefore real asset returns in general, in the economy. The real interest rate is essentially the real return that consumers/savers demand in exchange for postponing consumption. Faster growth in potential GDP means that consumers expect their real income to rise more rapidly. This implies that an extra unit of future income/consumption is less valuable than it would be if income were expected to grow more slowly. Hence, all else equal, the real interest rate will have to be higher in order to induce the savings needed to fund required capital accumulation. Thus, higher rates of potential GDP growth translate into higher real interest rates and higher expected real asset returns in general.

[8]Note that this is an argument about cyclical variations in growth and inflation around the economy's long-term potential growth rate. It does not imply that there is a long-run trade-off between growth and inflation.

Potential GDP and its growth rate enter into fixed-income analysis in other ways as well. Among them are the following:

- A higher rate of potential GDP growth improves the general credit quality of fixed-income securities because most such securities are ultimately backed by a flow of income even if the lender has a claim on specific underlying assets.
- Central banks frequently explain their monetary policy decisions by referring to the level of resource utilization and the degree of slack in the economy. In other words, monetary policy decisions are affected by the difference between an economy's estimated potential output and its actual operating level (referred to as the output gap) and by growth of actual GDP relative to the sustainable growth rate. Thus, fixed-income investors need to closely monitor the output gap and growth rates of actual and potential GDP to assess the likelihood of a change in central bank policy.
- Credit rating agencies use the growth rate of potential GDP as an input in evaluating the credit risk of sovereign debt or government-issued debt. All else held equal, slower estimated potential GDP growth raises the perceived risk of these bonds.
- Government budget deficits typically increase during recessions and decrease during expansions. In examining fiscal policy, actual fiscal positions are often judged relative to structural or cyclically adjusted deficits—the budgetary balance that would exist if the economy were operating at potential GDP.

EXAMPLE 11-2 Impact on Equity and Fixed-Income Investors

Your firm subscribes to asset class risk and return estimates generated by a large pension consultant. The equity market return estimates are based primarily on long-term average index returns. Following a multiyear period of very high equity returns driven by unusually high earnings growth and expanding P/E multiples, capital's share of total income as well as valuation multiples are near all-time highs. Based on the latest data, the vendor projects that your domestic equity market will return 13.5 percent per year—11 percent annual appreciation and 2.5 percent dividend yield—forever.

Your firm also subscribes to a macroeconomic forecasting service that provides, in addition to shorter-term projections, estimates of the long-term growth rate of potential GDP and the long-term inflation rate. This service forecasts 3.25 percent real growth in the future and 3.75 percent inflation, down from 4.0 percent and 5.0 percent, respectively, over the past 75 years.

1. Why might you have greater confidence in the macroeconomic service's forecasts than in the pension consultant's equity market return forecast?
2. Assuming the macroeconomic forecasts are accurate, what implicit assumptions underlie the pension consultant's forecast of 11 percent equity market appreciation?
3. Assuming the macroeconomic forecasts are accurate, what would be a more reasonable forecast for long-term equity returns?
4. In addition to its long-term potential GDP forecast, the macroeconomic forecasting service estimates sluggish 1.5 percent GDP growth for the next year. Based on this short-term GDP forecast, the bond analyst at your firm recommends that the firm increase its fixed-income investments. What assumptions underlie the bond analyst's forecast?

Solution to 1: High volatility makes equity returns very hard to predict based on their own history. As illustrated in Exhibit 11-3, the high volatility of equity returns is due to the underlying volatility of earnings as a share of GDP and valuation ratios. Long-term real GDP growth rates tend to be far less volatile, especially for developed economies, such as the United States or the euro area, because long-term potential growth is governed by fundamental economic forces that tend to evolve slowly over time. Similarly, for countries with prudent monetary policies, inflation rates are much less volatile than stock prices. Thus, one could reasonably place much higher confidence in forecasts of long-term real and nominal (real growth plus inflation) GDP growth than in equity market return forecasts based on historical equity returns.

Solution to 2: We can decompose the equity market appreciation rate into components due to (1) nominal GDP growth, (2) expansion/contraction of the share of profits in GDP, and (3) expansion/contraction of the P/E. The macroeconomic forecast indicates that nominal GDP will grow at 7 percent (3.25 percent real + 3.75 percent inflation). So the pension consultant's forecast of 11 percent equity market appreciation implies a 4 percent per year combined contribution from expansion in the P/E multiple and/or the profit share of GDP—forever.

Solution to 3: Neither the P/E nor the profit share of GDP can grow at a nonnegligible rate forever. A much more reasonable forecast of long-term equity market appreciation would be the projected 7 percent growth rate of nominal GDP.

Solution to 4: With forecasted actual GDP growth well below the growth in potential GDP, the bond analyst assumes a growing output gap or slack in the economy. This slack may place downward pressure on inflation and reduce inflationary expectations. To close this gap, the central bank may need to lower short-term interest rates and ease policy. In such an environment, bond prices should rise.

4. DETERMINANTS OF ECONOMIC GROWTH

What are the forces driving long-run economic growth? The following sections discuss labor, physical and human capital, technology, and other factors, such as natural resources and public infrastructure, as inputs to economic growth and production functions, and how changes in such inputs affect growth. Section 4.1 begins the discussion by presenting one of the simplest useful models of the production function.

4.1. Production Function

A production function is a model of the quantitative link between the inputs (factors of production), technology, and output. A two-factor aggregate production function with labor and capital as the inputs can be represented as:

$$Y = AF(K, L) \tag{11-1}$$

where Y denotes the level of aggregate output in the economy, L is the quantity of labor or number of workers or hours worked in the economy, and K is an estimate of the capital services provided

by the stock of equipment and structures used to produce goods and services. The function $F(\ )$ embodies the fact that capital and labor can be used in various combinations to produce output.

In this production function, A is a multiplicative scale factor referred to as **total factor productivity (TFP)**. Note that an increase in TFP implies a proportionate increase in output for any combination of inputs. Hence, TFP reflects the general level of productivity or technology in the economy. The state of technology embodies the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices.

It is worth noting that both the function $F(\ )$ and the scale factor A reflect technology. An innovation that makes it possible to produce the same output with the same amount of capital but fewer workers would be reflected in a change in the function $F(\ )$ because the relative productivity of labor and capital has been altered. In contrast, an increase in TFP does not affect the relative productivity of the inputs. As is standard in the analysis of economic growth, *unless stated otherwise, the level of technology should be interpreted as referring to TFP*.

In order to obtain concrete results, it is useful to use a specific functional form for the production function. The **Cobb–Douglas production function**, given by:

$$F(K, L) = K^{\alpha}L^{1-\alpha} \tag{11-2}$$

is widely used because it is easy to analyze and does a good job of fitting the historic data relating inputs and output. The parameter α determines the shares of output (factor shares) paid by companies to capital and labor and is assumed to have a value between 0 and 1. The reason for this follows from basic microeconomics. In a competitive economy, factors of production are paid their marginal product. Profit maximization requires that the marginal product of capital equal the **rental price of capital** and the marginal product of labor equal the (real) wage rate. In the case of capital, the marginal product of capital (MPK) for the Cobb–Douglas production function is:[9]

$$MPK = \alpha AK^{\alpha-1}L^{1-\alpha} = \alpha Y/K$$

Setting the MPK equal to the rental price (r) of capital,

$$\alpha Y/K = r$$

If we solve this equation for α, we find that it equals the ratio of capital income, rK to output or GDP, Y. Thus, α is *the share of GDP paid out to the suppliers of capital*. A similar calculation shows that $1 - \alpha$ is the share of income paid to labor. This result is important because it is easy to estimate α for an economy by simply looking at capital's share of income in the national income accounts.

The Cobb–Douglas production function exhibits two important properties that explain the relationship between the inputs and the output. First, the Cobb–Douglas production function exhibits **constant returns to scale**. This means that if all the inputs into the production process are increased by the same percentage, then output rises by that percentage.

[9]The marginal product of capital is simply the derivative of output with respect to capital. This can be approximated as $\Delta Y/\Delta K \approx [A(K + \Delta K)^{\alpha}L^{1-\alpha} - AK^{\alpha}L^{1-\alpha}]/\Delta K \approx (A\alpha K^{\alpha-1}\Delta KL^{1-\alpha})/\Delta K = A\alpha K^{\alpha-1}L^{1-\alpha} = \alpha Y/K$. The approximation becomes exact for very small increments, ΔK.

Under the assumption of constant returns to scale, we can modify the production function (Equation 11-1) and examine the determinants of the quantity of output per worker. Multiplying the production function by $1/L$ gives:

$$Y/L = AF(K/L, L/L) = AF(K/L, 1)$$

Defining $y = Y/L$ as the output per worker or (average) **labor productivity** and $k = K/L$ as the capital-to-labor ratio, this expression becomes:

$$y = AF(k, 1)$$

Specifying the Cobb–Douglas production function in output per worker terms, where again lowercase letters denote variables measured on a per capita basis, we get:

$$y = Y/L = A(K/L)^{\alpha}(L/L)^{1-\alpha} = Ak^{\alpha} \tag{11-3}$$

This equation tells us that the amount of goods a worker can produce (labor productivity) depends on the amount of capital available for each worker (capital-to-labor ratio), technology or TFP, and the share of capital in GDP (α). It is important to note that there are two different measures of productivity or efficiency in this equation. Labor productivity measures the output produced by a unit of labor and is measured by dividing the output (GDP) by the labor input used to produce that output ($y = Y/L$). TFP is a scale factor that multiplies the impact of the capital and labor inputs. Changes in TFP are estimated using a growth accounting method discussed in the next section.

A second important property of the model is the relationship between an individual input and the level of output produced. The Cobb–Douglas production function exhibits **diminishing marginal productivity** with respect to each individual input. Marginal productivity is the extra output produced from a one-unit increase in an input, keeping the other inputs unchanged. It applies to any input as long as the other inputs are held constant. For example, if we have a factory of a fixed size and we add more workers to the factory, the marginal productivity of labor measures how much additional output each additional worker will produce. Diminishing marginal productivity means that at some point the extra output obtained from each additional unit of the input will decline. To continue our example, if we hire more workers at the existing factory (fixed capital input in this case) each additional worker adds less to output than the previously hired worker does, and average labor productivity (y) falls.

The significance of diminishing marginal returns in the Cobb–Douglas production function depends on the value of α. A value of α close to zero means diminishing marginal returns to capital are very significant and the extra output made possible by additional capital declines quickly as capital increases. In contrast, a value of α close to 1 means that the next unit of capital increases output almost as much as the previous unit of capital. In this case, diminishing marginal returns still occur but the impact is relatively small. Note that the exponents on the K and L variables in the Cobb–Douglas production function sum to 1, indicating constant returns to scale; that is, there are no diminishing marginal returns if both inputs are increased proportionately.

4.2. Capital Deepening versus Technological Progress

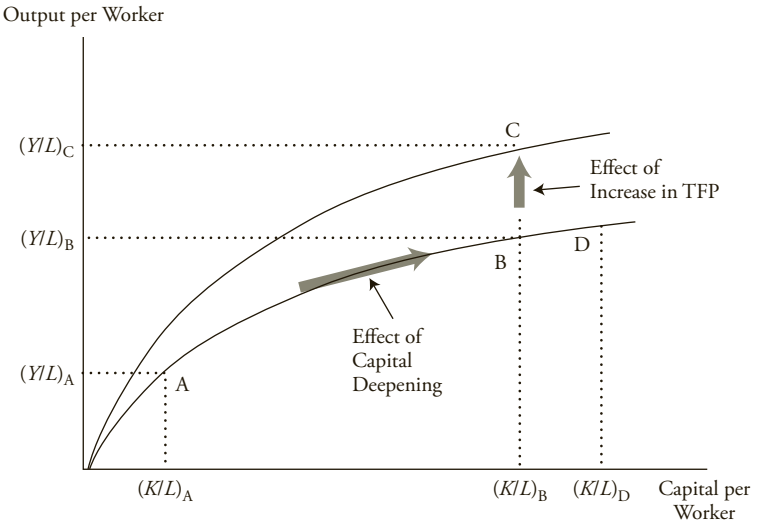
The property of diminishing marginal returns plays an important role in assessing the contribution of capital and technology to economic growth. Exhibit 11-4 shows the relationship

between per capita output and the capital-to-labor ratio. It shows that adding more and more capital to a fixed number of workers increases per capita output but at a decreasing rate. Looking at Equation 11-3 and Exhibit 11-4, we can think of growth in per capita output coming from two sources: capital deepening and an improvement in technology, often referred to as technological progress.

Capital deepening, an increase in the capital-to-labor ratio, is reflected in the exhibit by a move along the production function from point A to point B. The increase in the capital-to-labor ratio reflects rising investment in the economy. The ratio will increase as long as the growth rate of capital (net investment) exceeds the growth rate of labor. However, once the capital-to-labor ratio becomes very high, as at point B, further additions to capital have relatively little impact on per capita output (e.g., moving to point D). This occurs because the marginal product of capital declines as more capital is added to the labor input.

At the point where the marginal product of capital equals its marginal cost, profit maximizing producers will stop adding capital (i.e., stop increasing the capital-to-labor ratio).[10] As we will discuss in Section 5, this point is very significant in the neoclassical model of growth because per capita growth in the economy will come to a halt. Once the economy reaches this steady state, capital deepening cannot be a source of sustained growth in the economy. Only when the economy is operating below the steady state and when the marginal product of capital exceeds its marginal cost can capital deepening raise per capita growth.

EXHIBIT 11-4 Per Capita Production Function Capital Deepening versus Technological (TFP) Progress



[10]To avoid confusion later, we must note that once technological progress (TFP growth) is introduced, the capital-to-labor ratio will have to keep increasing just to keep the marginal productivity of capital equal to its marginal cost. But the point remains: Once that equality is attained, companies will not increase the capital-to-labor ratio faster than is necessary to maintain that equality.

The neoclassical model's stark implication that more rapid capital accumulation—that is, higher rates of investment—cannot result in a permanently higher rate of per capita growth is somewhat disappointing. As we will see in our discussion of endogenous growth, capital accumulation can result in a permanently higher growth rate if the investment results not just in *more* capital (i.e., pure capital deepening) but also in new, innovative products and processes. That is, if the additional capital embodies new, more efficient methods of production or previously unavailable products, then more rapid capital accumulation can result in a permanently higher growth rate of per capita output.

In contrast to moves along a given production function, an improvement in TFP causes a proportional upward shift in the entire production function. As a result, the economy can produce higher output per worker for a given level of capital per worker. This is shown in Exhibit 11-4 by the move from point B to point C. Technological progress also increases the marginal product of capital relative to its marginal cost. This makes additional capital investments profitable and tends to mitigate the limits imposed on growth by diminishing marginal returns. In addition, continued growth in per capita output is possible even in the steady state as long as there is ongoing technological progress (increases in TFP). In summary, *sustained growth in per capita output requires progress in TFP*.

EXAMPLE 11-3 Capital Deepening versus Technological Progress

One of main differences between developed and developing countries is the amount of capital available for each worker. Country A is an advanced economy with $100,000 of capital available for each worker and thus a high capital-to-labor ratio. In contrast, Country B is a developing country with only $5,000 of capital available for each worker. What impact will the following developments have on the growth rate of potential GDP?

1. An increase in business investment in both countries.
2. An increase in the amount of spending on university research in both countries.
3. An elimination of restrictions in Country B on the inflow of foreign investment.

Solution to 1: An increase in business investment will raise the capital-to-labor ratio in both countries. It results in capital deepening and a movement along the per worker production function. However, the impact on growth will be significantly different for the two countries. Country B will experience an increase in output per worker and thus in the growth rate of potential GDP. This is because Country B operates at a low level of capital per worker, at a point like A in Exhibit 11-4. Diminishing returns to capital are small, so any addition to capital has a major impact on growth. Country A operates at a point like B in Exhibit 11-4, so additions to capital have little impact on growth because of diminishing returns.

Solution to 2: An increase in spending on university research will increase TFP and cause an upward shift in the production function in both countries. This can be seen in the move from point B to point C in Exhibit 11-4. The shift in the production function will raise growth in both countries and offset the negative impact of diminishing returns. This result shows that developing countries have the potential to grow through

both capital deepening and technological progress, whereas improvement in potential GDP growth in developed countries is largely driven by technological progress.

Solution to 3: The elimination of restrictions will result in higher foreign investment, which has the same impact as an increase in domestic business investment. This is again a movement along the production function such as from point A to point B in Exhibit 11-4. With diminishing returns insignificant at low levels of capital to labor, the higher level of foreign investment will boost growth of potential GDP in Country B.

4.3. Growth Accounting

Since the publication of Solow's seminal work in 1957,[11] growth accounting has been used to analyze the performance of economies. The growth accounting equation is essentially the production function written in the form of growth rates. It starts with the Cobb–Douglas production function and decomposes the percentage change in output into components attributable to capital, labor, and technology:

$$\Delta Y/Y = \Delta A/A + \alpha\Delta K/K + (1 - \alpha)\Delta L/L \tag{11-4}$$

The **growth accounting equation** states that the growth rate of output equals the rate of technological change plus α times the growth rate of capital plus $(1 - \alpha)$ times the growth rate of labor. Because a 1 percent increase in capital leads to an α% increase in output, α is the elasticity of output with respect to capital. Similarly, $(1 - \alpha)$ is the elasticity of output with respect to labor. Thus, in the Cobb–Douglas production function, the exponents α and $(1 - \alpha)$ play dual roles as both output elasticities and the shares of income paid to each factor. Note that the impact of any unspecified inputs (e.g., natural resources) is subsumed into the TFP component.

Data on output, capital, labor, and the elasticities of capital and labor are available for most developed countries. The rate of technological change is not directly measured and must therefore be estimated. The elasticities of capital and labor in the growth accounting equation are the relative shares of capital (α) and labor $(1 - \alpha)$ in national income and are estimated from the GDP accounts. For the United States, the relative shares of labor and capital are approximately 0.7 and 0.3, respectively. This means that an increase in the growth rate of labor will have a significantly larger impact—roughly double—on potential GDP growth than will an equivalent increase in the growth rate of capital, holding all else equal. For example, because capital's share in GDP in the U.S. economy is 0.3, a 1 percent increase in the amount of capital available for each worker increases output by only 0.3 percent. An equivalent increase in the labor input would boost growth by 0.7 percent.

The growth accounting equation has a number of uses in studying an economy. First, Solow used the equation to estimate the contribution of technological progress to economic growth. Solow estimated the growth in TFP as a residual in the preceding equation by plugging in $\Delta Y/Y$, $\Delta K/K$, $\Delta L/L$, and α and solving for $\Delta A/A$. This residual measures the amount of output that cannot be explained by growth in capital or labor and can thus be regarded as progress in TFP.

[11] See Solow (1957).

Second, the growth accounting equation is used to empirically measure the sources of growth in an economy. In such studies, the growth accounting equation is used to quantify the contribution of each factor to long-term growth in an economy and answer such questions as the following: How important are labor and demographic factors to growth? What is the contribution of capital, and how important is capital deepening as a source of growth? What is the impact of TFP? The growth accounting equation can be expanded by considering different forms of capital and labor inputs, such as human capital and knowledge capital, and by considering the quality of the inputs as well.

Finally, the growth accounting equation is used to measure potential output. Potential GDP is estimated using Equation 11-4 with trend estimates of labor and capital and α estimated as 1 minus the labor share of GDP. The difficult task is estimating the growth rate of TFP, which, by definition, is a residual in the growth accounting equation.[12] The standard methodology treats TFP as exogenous and estimates its growth rate using various time-series models.

An alternative method of measuring potential GDP is the **labor productivity growth accounting equation**. It is very similar to the Solow approach but is simpler and models potential GDP as a function of the labor input and the productivity of the labor input. It avoids the need to estimate the capital input and the difficulty associated with computing total factor productivity. The disadvantage is that it incorporates both capital deepening and TFP progress in the productivity term in a way that can be difficult to analyze and to predict over long periods of time. Under this approach, the equation for estimating potential GDP is:

$$\begin{aligned}\text{Growth rate in potential GDP} = {} & \text{Long-term growth rate of labor force} \\ & + \text{Long-term growth rate in labor productivity} \quad (11\text{-}5)\end{aligned}$$

Thus, potential GDP growth is a combination of the long-term growth rate of the labor force and the long-term growth rate of labor productivity. If the labor force is growing at 1 percent per year and productivity per worker is rising at 2 percent per year, then potential GDP is rising at 3 percent per year.

4.4. Extending the Production Function

As a simplification, the production function in Equation 11-1 focused on only the labor and capital inputs. A more complete specification of the production function expands the list of inputs to include the following:

- Raw materials: natural resources such as oil, lumber, and available land (N).
- Quantity of labor: the number of workers in the country (L).
- Human capital: education and skill level of these workers (H).
- Information, computer, and telecommunications (ICT) capital: computer hardware, software, and communication equipment (K_{IT}).
- Non-ICT capital: transport equipment, metal products and plant machinery other than computer hardware and communications equipment, and nonresidential buildings and other structures (K_{NT}).

[12] TFP is computed as the growth in output less the growth in the factor inputs. These would include labor and capital in the traditional Solow two-factor production model. If the production function is expanded by including more inputs, the weighted growth rates of these inputs would also be subtracted from the growth in output.

- Public capital: infrastructure owned and provided by the government (K_P).
- Technological knowledge: the production methods used to convert inputs into final products, reflected by total factor productivity (A).

The expanded production function is expressed mathematically as:

$$Y = AF(N, L, H, K_{IT}, K_{NT}, K_p)$$

The impact of each of these inputs on economic growth is addressed in the following sections.

4.5. Natural Resources

Raw materials, including everything from available land to oil to water, are an essential input to growth. There are two categories of natural resources:

1. **Renewable resources** are those that are replenished, such as forests. For example, if a tree is cut, a seedling can be planted and a new forest can be harvested in the future.
2. **Nonrenewable resources** are finite resources that are depleted once they are consumed. Oil and coal are examples.

Although it seems intuitive that countries with more natural resources will be wealthier, the relationship between resource endowment and growth is not so straightforward. Natural resources do account for some of the differences in growth among countries. Today, Middle Eastern countries and such countries as Brazil and Australia have relatively high per capita incomes because of their resource base. Countries in the Middle East have large pools of oil. Brazil has an abundance of land suitable for large-scale agricultural production, allowing it to be a major exporter of coffee, soybeans, and beef.

Even though *access* to natural resources (e.g., via trade) is important, *ownership and production of natural resources is not necessary for a country to achieve a high level of income*. Countries in East Asia, such as Japan and South Korea, have experienced rapid economic growth but have few natural resources. In contrast, both Venezuela and Saudi Arabia have large oil reserves and are major producers of oil, yet both countries have experienced subpar growth in comparison to the natural-resource-poor countries of Singapore, Japan, and South Korea. As was examined in Example 11-1, economic growth in Venezuela over the past 60 years was well below economic growth of Singapore, Japan, and South Korea.

For some countries, the presence of natural resources may even restrain growth, resulting in a so-called resource curse. Venezuela and Nigeria are two examples of countries blessed with resources yet with sluggish economic growth. There are two main reasons why this may occur. First, countries rich in natural resources may fail to develop the economic institutions necessary for growth. Second, countries rich in resources may suffer the **Dutch disease**, where currency appreciation driven by strong export demand for resources makes other segments of the economy, in particular manufacturing, globally uncompetitive.[13] In this situation, the manufacturing sector contracts and the country does not participate in the TFP progress that occurs in countries with more vigorous manufacturing sectors.

[13]Following the discovery of large natural gas fields in the Netherlands, the Dutch guilder appreciated and the manufacturing sector contracted.

In contrast, there is a long-standing concern that nonrenewable natural resources will eventually limit growth. The idea is that a combination of rapid economic growth and a fixed stock of resources will cause resource depletion as the available pool of resources is used up. These concerns are probably overstated. Technological progress (TFP from all sources) enables the economy to use fewer resources per unit of output and to develop substitutes. The growing scarcity of specific resources will increase their prices and encourage a shift toward more plentiful substitutes. Finally, the share of national income going to land and resources has been declining for most countries, especially as the composition of output in the global economy shifts toward the use of more services.

EXAMPLE 11-4 Impact of Natural Resources

1. The following table shows the share of world proved oil reserves as of 1990 for each of the 34 countries shown in Exhibit 11-1, along with the growth rate of real per capita GDP from 1990 to 2010. The simple correlation between the share of oil reserves and subsequent growth is not statistically different from zero.

	Percent of World Proved Oil Reserves, 1990	Real Per Capita GDP Growth (%), 1990–2010
Saudi Arabia	25.75	3.00
Venezuela	5.85	2.80
Mexico	5.64	2.65
United States	2.62	2.50
China	2.40	10.45
Nigeria	1.60	5.25
Indonesia	0.82	4.60
India	0.75	6.55
Canada	0.61	2.10
Egypt	0.45	4.65
United Kingdom	0.43	2.15
Brazil	0.28	3.05
Argentina	0.23	4.40
Australia	0.17	3.20
Italy	0.07	1.00
Turkey	0.05	3.80

(*Continued*)

	Percent of World Proved Oil Reserves, 1990	Real Per Capita GDP Growth (%), 1990–2010
Peru	0.04	4.85
Germany	0.04	1.60
France	0.02	1.45
New Zealand	0.01	2.60
Pakistan	0.01	4.35
Japan	0.01	1.05
Spain	0.00	2.50
Philippines	0.00	3.85
Taiwan	0.00	5.14
Botswana	0.00	5.29
Ethiopia	0.00	5.61
Hong Kong	0.00	4.00
Ireland	0.00	4.83
Kenya	0.00	2.89
Singapore	0.00	6.45
South Africa	0.00	2.65
South Korea	0.00	5.69
Vietnam	0.00	7.45

Sources: U.S. Energy Information Administration (www.eia.gov) and Exhibit 11-1.

What might account for the fact that real per capita GDP growth appears to be unrelated to oil reserves, perhaps the single most economically important natural resource (aside from water)?

Solution: Energy is a vital input for any economy. Thus, *access* to energy resources is critical. *Ownership* of raw energy resources, however, is not. Countries that are not self-sufficient in oil or other resources acquire what they need through trade. It should be noted that countries that lack oil may possess other types of energy resources, such as natural gas, coal, hydropower, or geothermal energy. In addition, countries can grow by emphasizing less energy-intensive products, especially services, and adopting more energy-efficient production methods. In sum, natural resources are important but not necessary for growth.

4.6. Labor Supply

As noted earlier, economic growth is affected by increases in inputs, mainly labor and capital. Growth in the number of people available for work (quantity of workforce) is an important source of economic growth and partially accounts for the superior growth performance of the United States among the advanced economies—in particular, relative to Europe and Japan. Most developing countries, such as China, India, and Mexico, have a large potential labor supply. We can measure the potential size of the labor input as the total number of hours available for work. This, in turn, equals the labor force times the average hours worked per worker. The **labor force** is defined as the working age population (ages 16 to 64) that is either employed or available for work but not working (i.e., unemployed). Thus, growth in the labor input depends on four factors: population growth, labor force participation, net migration, and average hours worked.

4.6.1. Population Growth

Long-term projections of the labor supply are largely determined by the growth of the working age population. Population growth is determined by fertility rates and mortality rates. Population growth rates are significantly lower in the developed countries than in the developing countries. (See Exhibit 11-5.) As a result, there is an ongoing decline in the developed countries' share of the world's population. Note that although population growth may increase the growth rate of the overall economy, it has no impact on the rate of increase in *per capita* GDP.

The age mix of the population is also important. The percentage of the population over the age of 65 and the percentage below the age of 16 are key considerations. Some of the developed countries, especially European countries, Japan, and South Korea, are facing a

EXHIBIT 11-5 Population Data for Selected Countries (millions)

	2000	2005	2010	Annual Growth (%), 2000–2010
France	59.1	61.2	63.0	0.64
Germany	82.2	82.3	81.6	−0.07
Ireland	3.8	4.0	4.5	1.71
Spain	40.3	43.4	46.1	1.35
United Kingdom	58.9	59.4	61.3	0.40
Russia	146.7	142.8	142.9	−0.26
Japan	126.9	127.8	127.6	0.06
United States	282.2	295.6	309.1	0.91
Mexico	98.4	103.9	108.4	0.97
China	1,267.4	1307.6	1341.4	0.57
India	1,024.3	1,110.0	1,190.5	1.52

Source: OECD.StatExtracts.

growing demographic burden as the portion of nonworking elders (over 65) grows as a share of the population. In contrast, growth in many developing countries will receive a demographic boost as the fraction of the population below the age of 16 begins to decline. Interestingly, China is similar to the advanced economies, with a growing proportion of the population over age 65.

4.6.2. Labor Force Participation

In the short run, the growth rate of the labor force may differ from population growth because of changes in the participation rate. The **labor force participation rate** is defined as the percentage of the working age population in the labor force. It has trended upward in most countries over the past few decades because of rising participation rates among women. In contrast to population, an increase in the participation rate may raise the growth of per capita GDP. In many southern European countries, such as Greece and Italy, the participation rate among women is well below the rates in the United States and northern European countries (see Exhibit 11-6). Thus, rising participation rates among women in these countries could increase growth in the labor force and in potential GDP. This has been the case for Spain, where the female labor force participation rate rose from 52.0 percent in 2000 to 66.1 percent in 2010. It should be noted, however, that rising or falling labor force participation is likely to represent a transition to a new higher or lower level of participation rather than a truly permanent rate of change. Thus, although trends in participation may contribute to or detract from potential growth for substantial periods, one should be cautious in extrapolating such trends indefinitely.

EXHIBIT 11-6 Labor Force Data for Selected Countries, 2010

	Percentage of Population under Age 15	Percentage of Population over Age 65	Participation Rate: Male (%)	Participation Rate: Female (%)
France	18.3	16.7	72.6	67.3
Germany	13.1	21.0	82.3	71.5
Greece	14.3	18.6	76.5	55.0
Ireland	21.5	11.4	79.5	63.4
Italy	14.1	20.0	73.5	51.5
Spain	15.0	17.0	80.4	66.1
Sweden	16.6	18.3	84.5	77.8
United Kingdom	17.7	15.9	82.8	70.5
Japan	13.3	22.7	93.8	68.5
United States	20.1	13.1	79.2	69.8
Mexico	28.1	5.9	83.2	48.3
Turkey	26.4	7.0	73.9	28.8

Source: OECD.StatExtracts.

EXAMPLE 11-5 Impact of the Age Distribution on Growth: Mexico versus Germany

Exhibits 11-5 and 11-6 provide population data for selected countries. The data show that the rate of population growth and the age composition vary significantly among countries. Thus, demographic factors can be expected to have a significant impact on relative growth rates across countries. This is very clear in the cases of Mexico and Germany. There was essentially zero growth in the population of Germany from 2000 to 2010, while the population of Mexico increased by 0.97 percent annually. The age composition of the two countries is also very different. How will the age distribution impact growth over the next decade?

Solution: What is important for growth is the number of workers available to enter the workforce. Over the next decade, Mexico will receive a demographic benefit because of the high percentage of young people entering the workforce. This is because 28.1 percent of the population in 2010 was below the age of 15. In contrast, only 13.1 percent of the German population was below the age of 15. In addition, Germany is facing a demographic challenge given the high and growing share of its population over the age of 65. In Mexico, only 5.9 percent of the population is above the age of 65, compared with 21 percent in Germany. In sum, the lack of population growth and a rapidly aging population in Germany will limit its potential rate of growth. Germany must rely on high labor productivity growth, increase its workforce participation rate, or encourage immigration if it is to increase its near-term potential rate of growth. Meanwhile, potential GDP growth in Mexico should receive a boost from its favorable population trends.

4.6.3. Net Migration

Another factor increasing economic and population growth, especially among the developed countries, is immigration. Heightened immigration is a possible solution to the slowing labor force growth being experienced by many developed countries with low birthrates within the native population. The growth rate of the labor force in Ireland, Spain, the United Kingdom, and the United States has increased over the past decade because of immigration. As Exhibit 11-5 shows, the population growth rates for Ireland and Spain for the period 2000–2010, at 1.71 percent and 1.35 percent, respectively, were well above the population growth rates in other European countries. As shown in Exhibit 11-7, this is due to the impact of immigration. The open-border policies of both countries led to a significant population of immigrants that contributed to a large increase in labor input for both countries. As a consequence, both countries experienced GDP growth above the European average during this period (see Exhibit 11-1).

EXHIBIT 11-7 Ireland and Spain: Net Migration

	2000–2007	2008	2009	2010	Total 2000–2010
Ireland	357,085	38,502	−7,800	−12,200	375,587
Spain	4,222,813	460,221	181,073	111,249	4,975,356

Source: OECD.StatExtracts.

EXAMPLE 11-6 Potential Growth in Spain: Labor Input

The investment policy committee of Global Invest Inc. reviewed a report on the growth prospects for Spain and noted that, with total hours worked growing at a 1.2 percent annual rate between 2000 and 2010, labor input had been a major source of growth for the economy. As of 2011, some members expected the growth rate of labor to slow considerably given projection from the OECD and IMF that immigration into Spain will fall to essentially zero over the next few years. A research assistant at the firm gathered demographic data on Spain from Exhibits 11-5, 11-6, and 11-7 and other sources. The data are presented in the following table:

	2000	2010	Annual Growth (2000–2010)
Population (millions)	40.3	46.1	1.35%
Immigration since 2000 (millions)		4.975	
Percent of population under 15		15.0	
Percent of population over 65		17.0	
Male participation rate		80.4%	
Female participation rate		66.1%	
Unemployment rate		20.1%	

Using this information for Spain and Exhibits 11-5 and 11-6 for relevant comparison data, determine the following:

1. Whether a change in the trend growth rate of the labor input is likely over the next few years.
2. How the high unemployment rate of 20.1 percent is likely to affect the growth rate of the labor force.

Solution to 1: The growth in the labor input depends on a number of factors, including the population growth rate, the labor force participation rate, and the percentage of the population below the age of 15. The labor force in Spain expanded sharply between 2000 and 2010 mainly because of a large 5.8 million person increase in the population, going from 40.3 million in 2000 to 46.1 million in 2010. Looking ahead, growth in the labor force is set to slow substantially for a number of reasons:

- The population increase between 2000 and 2010 is very misleading and not likely to be repeated in the future. Between 2000 and 2010, immigration raised the population of Spain by nearly 5 million people. Without the immigrants, the population of Spain between 2000 and 2010 would have grown by only about 825,000 or at an annual rate of 0.2 percent. With immigration, the population growth rate was 1.35 percent. The pace of immigration that occurred between 2000 and 2010 is not sustainable and is likely to slow. This will result in slower growth in the population and the labor force.

- In the short run, the growth rate of the labor force may differ from population growth because of changes in the participation rate. Looking at the data, the male participation rate in Spain, at 80.4 percent, is very high and, as shown in Exhibit 11-6, is above the male participation rates in France, Greece, and Italy and slightly below that of Germany. The female participation rate is low in comparison to northern European countries, such as Sweden. But it is higher than in Greece and Italy, which are probably better comparisons. Thus, little increase is likely in the male or female participation rates.
- Only 15 percent of the Spanish population is below the age of 15. The comparable figure from Exhibit 11-6 for the United Kingdom is 17.7 percent, for France 18.3 percent, for the United States 20.1 percent, and for Mexico 28.1 percent. Thus, Spain does not appear poised for a notable surge in young adults entering the labor force.

In summary, growth in the labor input in Spain should slow over the next few years, and the growth rate of potential GDP should do the same.

Solution to 2: Reducing the unemployment rate would mitigate some of the negative demographic factors because a reduction in the number of unemployed workers would boost utilization of the existing labor supply. This would represent a transition to a higher level of employment rather than a permanent increase in the potential growth rate. Nonetheless, it could boost potential growth for a substantial period.

4.6.4. Average Hours Worked

The contribution of labor to overall output is also affected by changes in the average hours worked per worker. Average hours worked is highly sensitive to the business cycle. However, the long-term trend in average hours worked has been toward a shorter workweek in the advanced countries. This development is the result of legislation, collective bargaining agreements, the growth of part-time and temporary work, and the impact of both the wealth effect and high tax rates on labor income, which cause workers in high-income countries to value a certain amount of leisure time relatively highly compared to the forgone labor income.

Exhibit 11-8 provides data on average hours worked per year per person in the labor force for selected years since 1995. For most countries, the average number of hours worked per year has been declining. There is also a significant difference in hours worked across countries. In 2010, average hours worked per year in South Korea, at 2,193 hours, were 54.5 percent more than the 1,419 average hours worked per year in Germany. The increase in female labor force participation rates may be contributing to the shorter average workweek because female workers disproportionately take on part-time, rather than full-time, jobs.

4.7. Labor Quality: Human Capital

In addition to the quantity of labor, the quality of the labor force is an important source of growth for an economy. **Human capital** is the accumulated knowledge and skills that workers acquire from education, training, or life experience. In general, better-educated and more skilled workers will be more productive and more adaptable to changes in technology or other shifts in market demand and supply.

EXHIBIT 11-8 Average Hours Worked per Year per Person in Selected Countries

	1995	2000	2005	2010
France	1,651	1,591	1,559	1,594
Germany	1,534	1,473	1,435	1,419
Greece	2,123	2,121	2,081	2,109
Ireland	1,875	1,719	1,654	1,664
Italy	1,859	1,861	1,819	1,778
Spain	1,733	1,731	1,688	1,663
Sweden	1,609	1,574	1,607	1,624
United Kingdom	1,743	1,711	1,676	1,647
Japan	1,884	1,821	1,775	1,733
South Korea	2,658	2,520	2,364	2,193
Canada	1,761	1,768	1,738	1,702
United States	1,840	1,832	1,795	1,778
Mexico	1,857	1,888	1,909	1,866
Turkey	1,876	1,937	1,918	1,877

Source: OECD.StatExtracts.

An economy's human capital is increased through investment in education and on-the-job training. Like physical capital, investment in education is costly, but studies show that there is a significant return on that investment. That is, people with more education earn higher wages. In addition, education may also have a spillover or externality impact. Increasing the educational level of one person raises not only the output of that person but also the output of those around that person. The spillover effect operates through the link between education and advances in technology. Education not only improves the quality of the labor force, and thus the stock of human capital, but also encourages growth through innovation. Importantly, increased education, obtained both formally and via on-the-job training, could result in a permanent increase in the growth rate of an economy if the more educated workforce results in more innovations and a faster rate of technological progress. Investment in the health of the population is also a major contributor to human capital, especially in the developing countries.

4.8. Capital: ICT and Non-ICT

The physical capital stock increases from year to year as long as net investment (gross investment less the depreciation of the capital) is positive. Thus, countries with a higher rate of investment should have a growing physical capital stock and a higher rate of GDP growth.[14]

[14]The impact on growth of per capita GDP will be somewhat smaller if the population is growing, because a proportion of net investment simply provides the capital needed to maintain the capital-to-labor ratio.

EXHIBIT 11-9 Business Investment as a Percentage of GDP

	ICT Percentage of GDP			Investment Percentage of GDP			
	1990	2000	2008	1990	2000	2008	2010
Developed Countries							
France	2.6	3.7	3.2	21.5	19.5	21.1	19.1
Germany	3.1	3.6	2.7	22.8	21.5	19.4	17.3
Ireland	1.1	2.3	1.2	20.8	23.9	21.7	11.0
Italy	2.4	2.8	2.1	22.0	20.3	21.1	20.2
Spain	3.4	3.6	3.7	25.3	26.2	28.8	23.0
United Kingdom	3.3	5.1	4.2	20.5	17.1	16.6	15.0
Australia	3.6	5.5	4.6	23.6	22.0	29.4	27.6
Japan	4.9	3.7	3.1	32.5	25.4	23.2	20.2
South Korea	2.4	5.1	4.8	35.7	30.6	31.2	29.1
Singapore	3.2	5.4	4.9	35.2	33.1	30.2	23.8
Canada	2.8	3.9	3.6	21.3	19.2	22.6	22.2
United States	4.1	6.6	5.1	17.4	19.9	18.1	15.8
Developing Countries							
Brazil	NA	NA	NA	14.0	18.3	20.7	19.3
China	NA	NA	NA	24.9	35.1	44.0	48.2
India	NA	NA	NA	21.8	24.3	34.9	36.8
Mexico	NA	NA	NA	17.9	25.5	26.9	25.0
South Africa	NA	NA	NA	19.1	15.1	22.5	19.3

Source: OECD StatLink.

Exhibit 11-9 shows the level of gross nonresidential investment as a share of GDP. The exhibit shows significant variation across countries, with the investment share in the United States being low in comparison to other developed countries.

The correlation between economic growth and investment is high. Countries that devote a large share of GDP to investment, such as China, India, and South Korea, have high growth rates. The fastest-growing countries in Europe over the past decade, Ireland and Spain, have the highest investment-to-GDP ratios. Countries that devote a smaller share of GDP to investment, such as Brazil and Mexico, have slower growth rates. The data show why the Chinese economy has expanded at such a rapid rate: annual GDP growth rate in excess of 10 percent over the past two decades. Investment spending in China on new factories, equipment, and infrastructure as a percentage of GDP is the highest in the world. In recent years, China devoted over 40 percent of its GDP to investment spending.

As we discussed in Section 4.2, long-term sustainable growth cannot rely on pure capital deepening. How can we reconcile this notion with the strong correlation between investment

spending and economic growth across countries? First, although diminishing marginal productivity will eventually limit the impact of capital deepening, investment-driven growth may last for a considerable period of time, especially in countries that start with relatively low levels of capital per worker.

A second, and closely related, explanation is that the impact of investment spending on available capital depends on the existing physical capital stock. As with the share of GDP devoted to investment, the stock of capital available per worker varies significantly across countries. In 2000, the average U.S. worker had $148,091 worth of capital, compared with $42,991 in Mexico and $6,270 in India.[15] The wide difference in physical capital per worker suggests that the positive impact of changes in the physical capital stock on growth is very significant in developing countries. Mexican workers have relatively little access to machinery or equipment, so adding even a little can make a big percentage difference. In developed countries, such as the United States, Japan, Germany, France, and the United Kingdom, the physical capital stock is so large that positive net investment in any given year has only a small percentage effect on the accumulated capital stock. For the developed countries, a sustained high level of investment over many years is required to have a meaningful relative impact on the physical capital stock even though the absolute size of the increase in any given year is still larger than in the developing countries.

Third, because physical capital is not really homogeneous, the composition of investment spending and the stock of physical capital matters for growth and productivity. Insights obtained from the endogenous theory of growth (discussed in Section 5) and from studies attempting to obtain a more accurate measure of TFP show that the composition of the physical capital stock is very important. These studies suggest that capital spending could be separated into two categories. The first is spending on information, computers, and telecommunications equipment (ICT investment). Capital spending on these goods is a measure of the impact of the information technology (IT) sector on economic growth. One of the key drivers of growth in the developed countries over the past decade has been the IT sector. Growth in the IT sector has been driven by technological innovation that has caused the price of key technologies, such as semiconductors, to fall dramatically. The steep decline in the price of high-technology capital goods has encouraged investment in IT at the expense of other assets.

The IT sector has grown very rapidly and has made a significant contribution to increasing the rate of economic and productivity growth. The greater use of IT equipment in various industries has resulted in **network externalities**. Computers allow people to interconnect through the Internet and by e-mail, enabling them to work more productively. *The more people in the network, the greater the potential productivity gains*. The effects of the network externalities are largely captured in TFP rather than observed as a distinct, direct effect. The share of ICT investment in GDP is shown in Exhibit 11-9. The data show that in most countries the IT sector is still relatively small, and that between 2000 and 2008 IT spending declined as a share of GDP as the global recession disproportionately affected high-technology spending.

The other category of investment, non-ICT capital spending, includes nonresidential construction, transport equipment, and machinery. High levels of capital spending for this category should eventually result in capital deepening and thus have less impact on potential

[15]Heston, Summers, and Aten (2009).

GDP growth. In contrast, a growing share of ICT investments in the economy, through their externality impacts, may actually boost the growth rate of potential GDP.[16]

4.9. Technology

The most important factor affecting growth of per capita GDP is technology, especially in developed countries. Technology allows an economy to overcome some of the limits imposed by diminishing marginal returns and results in an upward shift in the production function, as we noted in Exhibit 11-4. Technological progress makes it possible to produce more and higher-quality goods and services with the same resources or inputs. It also results in the creation of new goods and services. Technological progress can also be one of the factors improving how efficiently businesses are organized and managed.

Technological change can be embodied in human capital (knowledge, organization, information, and experience base) or in new machinery, equipment, and software. Therefore, high rates of investment are important, especially investment in ICT goods. Countries can also innovate through expenditures, both public and private, on research and development (R&D). Expenditures on R&D and the number of patents issued, although not directly measuring innovation, provide some useful insight into innovative performance. Exhibit 11-10 shows R&D spending as a share of GDP for various countries. The developed countries spend the highest percentage of GDP on R&D because they must rely on innovation and the development of new products and production methods for growth.[17] In contrast, developing countries spend less on R&D because these countries can acquire new technology through imitation or copying the technology developed elsewhere. The embodiment of technology in capital goods can enable relatively poor countries to narrow the gap relative to the technology leaders.

The state of technology, as reflected by total factor productivity, embodies the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices. Because it is measured as a residual, TFP estimates are very sensitive to the measurements of the labor and capital inputs. Recent empirical work at the Conference Board and the OECD accounts for changes in the composition and quality of both the labor and capital inputs. The resulting measure of TFP should capture the technological and organizational improvements that increase output for a given level of inputs. Exhibit 11-11 provides data for the periods 1995–2005 and 2005–2009 on the growth rate in labor productivity and total factor productivity.[18] Labor productivity growth depends on both capital deepening

[16]It is worthwhile to note that there have been important transformational technologies at various stages of history. One need only think about the impact of the steam engine, the internal combustion engine, powered flight, atomic energy, vaccination, and so on to realize that revolutionary advances are not unique to information, computers, and telecommunications. All of these are, to some extent, general-purpose technologies (GPTs) that affect production and/or innovation in many sectors of the economy. ICT capital clearly embodies this GPT characteristic. Nanotechnology could well become the next "super GPT," at which point investing in ICT may begin to look like mere capital deepening.

[17]The relationship between economic growth and R&D spending is not clear-cut. Although technological innovation resulting from high R&D spending raises output and productivity in the long run, it may result in a cyclical slowing of growth as companies and workers are displaced by the new technologies. This is the Schumpeterian concept of creative destruction, which captures the double-edged nature of technological innovation.

[18]Data for the developing countries are from 1995–2005 and 2005–2008. TFP data are not yet available for 2009 for these countries.

EXHIBIT 11-10 Research and Development as a Percentage of GDP in Selected Countries

	1990	2000	2009
France	2.3	2.2	2.2
Germany	2.6	2.5	2.8
Ireland	0.8	1.2	1.8
Italy	1.2	1.0	1.3
Spain	0.8	1.0	1.4
United Kingdom	2.1	1.8	1.9
Australia	1.3	1.5	2.2
Japan	3.0	3.0	3.4
South Korea	1.7	2.3	3.1
Singapore	1.1	1.9	2.9
Canada	1.5	1.9	2.0
United States	2.6	2.7	2.9
China	NA	1.0	1.7
India	NA	0.8	0.8
Mexico	NA	0.3	0.4

Source: OECD.StatExtracts.

EXHIBIT 11-11 Labor and Total Factor Productivity

	Growth in Hours Worked[a] (%)	Growth in Labor Productivity (%)	Growth in TFP (%)	Growth Due to Capital Deepening (%)	Growth in GDP (%)	Productivity Level 2010; GDP per Hour Worked ($)
Germany						53.6
1995–2005	−0.3	1.6	0.9	0.7	1.3	
2005–2009	0.2	0.2	0.1	0.1	0.4	
Ireland						50.3
1995–2005	3.2	4.1	1.7	2.4	7.3	
2005–2009	−0.8	0.8	−2.1	2.9	0.0	
United States						60.3
1995–2005	0.9	2.4	0.9	1.5	3.3	
2005–2009	−0.8	1.5	−0.5	2.0	0.7	

EXHIBIT 11-11 *Continued*

	Growth in Hours Worked[a] (%)	Growth in Labor Productivity (%)	Growth in TFP (%)	Growth Due to Capital Deepening (%)	Growth in GDP (%)	Productivity Level 2010; GDP per Hour Worked ($)
Japan						40.7
1995–2005	−1.0	2.1	0.4	1.7	1.1	
2005–2009	−1.3	0.8	−0.6	1.4	−0.5	
South Korea						27.9
1995–2005	0.0	4.3	2.4	1.9	4.3	
2005–2009	−0.5	2.8	2.0	0.8	2.3	
China						8.6
1995–2005	1.1	6.7	1.5	5.2	7.8	
2005–2008	1.2	10.3	4.2	6.1	11.5	
India						5.3
1995–2005	2.1	4.2	1.9	2.3	6.3	
2005–2008	2.2	6.0	2.4	3.6	8.2	
Brazil						10.4
1995–2005	2.1	0.3	−0.3	0.6	2.4	
2005–2008	2.0	2.9	−0.5	3.4	4.9	
Mexico						16.8
1995–2005	2.2	1.4	0.4	1.0	3.6	
2005–2008	1.8	0.8	−0.1	0.9	2.6	

[a]Total hours worked is the preferred measure of labor quantity. However, this measure is not available for most developing countries (including China, India, Brazil, and Mexico). For these countries, total employment is used, assuming that the change in total hours worked equals the change in employment. In this case, labor productivity is measured as output per worker, but for the developed countries labor productivity is output per hour.

Source: Conference Board Total Economy Database.

and technological progress. The contribution of capital deepening can be measured as the difference between the growth rates of labor productivity and total factor productivity. For example, from 2005 to 2009, Ireland's labor productivity grew by 0.8 percent per year, of which 2.9 percent [0.8% – (–2.1%)] came from capital deepening, which offset the −2.1 percent decline in TFP. The larger the difference between the productivity growth measures, the more important capital deepening is as a source of economic growth. As we discussed previously, however, growth in per capita income cannot be sustained perpetually by capital deepening.

Exhibit 11-11 also provides data on the *level* of labor productivity or the amount of GDP produced per hour of work. The level of productivity depends on the accumulated stock of human and physical capital and is much higher among the developed countries. For example,

China has a population of over 1.3 billion people, compared with slightly over 300 million people in the United States. Although the United States has significantly fewer workers than China because of its smaller population, its economy as measured by real GDP is much larger. This is because U.S. workers have historically been much more productive than Chinese workers. As shown in Exhibit 11-11, the United States has had the highest level of productivity in the world, producing over $60 of GDP per hour worked. In comparison, Chinese workers produce only $8.6 worth of GDP per hour worked. Thus, U.S. workers are seven times more productive than Chinese workers. In contrast to the *level* of productivity, the *growth rate* of productivity will typically be higher in the developing countries, where human and physical capital are scarce but growing rapidly and the impact of diminishing marginal returns is relatively small.

An understanding of productivity trends is critical for global investors. A permanent increase in the rate of labor productivity growth will increase the sustainable rate of economic growth and raise the upper boundary for earnings growth and the potential return on equities. In contrast, a low growth rate of labor productivity, if it persists over a number of years, suggests poor prospects for equity prices. A slowdown in productivity growth lowers both the long-run potential growth rate of the economy and the upper limit for earnings growth. Such a development would be associated with slow growth in profits and correspondingly low equity returns.

EXAMPLE 11-7 Why the Sluggish Growth in the Japanese Economy?

As shown in Exhibit 11-1, annual growth in real GDP in Japan averaged 0.8 percent for 2001–2010 and a weak 1.3 percent in the prior decade. This growth is in sharp contrast to the 4.2 percent annual growth rate experienced from 1971 to 1990. The sluggish growth in Japan over the past decade should not be surprising. The economy of Japan is growing at its potential rate of growth, which is limited by the following three factors:

1. The labor input is not growing. Population growth has been essentially zero since 2000 (Exhibit 11-5), and the average number of hours worked per year per person is declining (Exhibit 11-8).
2. There has been a lack of technological innovation. The lack of growth in the labor input could be offset through higher productivity derived from innovation and more efficient use of available inputs. However, this is not occurring in Japan. Total factor productivity (Exhibit 11-11) increased at a sluggish 0.4 percent annual rate from 1995 to 2005 and declined between 2005 and 2009.
3. Diminishing returns to capital are very significant. Despite the negative growth in TFP, labor productivity growth remained relatively high. This means that all the growth in labor productivity in Japan was due to capital deepening (Exhibit 11-11). The problem for Japan, as discussed in Section 4.2, is that once the capital-to-labor ratio becomes high, further additions to capital have little impact on per capita output. Thus, the growth in labor productivity should slow.

Use the data for 2005–2009 and the labor productivity growth accounting equation to estimate the growth rate in potential GDP for Japan.

Solution: To estimate the growth rate in potential GDP, we use Equation 11-5, given by:

$$\text{Growth rate of potential GDP} = \text{Long-term growth rate of labor force} + \text{Long-term growth rate in labor productivity}$$

To use this equation, we need to project the growth rate in the labor input and labor productivity.

The hours worked data in Exhibit 11-11 are a potential source to use to estimate the growth rate of the labor input. Exhibit 11-11 shows the labor input for Japan declining by 1.3 percent per year between 2005 and 2009. The problem here is that the decline in hours worked is overstated because of the negative impact of the global recession on hours worked. As an alternative, the labor input should grow at the same rate as the population plus the net change in immigration. The population data in Exhibit 11-5 show essentially zero population growth in Japan for the period 2000–2010. This trend is likely to continue. Thus, a reasonable estimate for potential GDP growth in Japan is around 0.8 percent. We get this estimate by assuming no growth in the labor input and a 0.8 percent annual increase in labor productivity (using data from Exhibit 11-11 for 2005–2009).

4.10. Public Infrastructure

The final expansion of the definition of the capital input is public infrastructure investment. Roads, bridges, municipal water, dams, and, in some countries, electric grids are all examples of public capital. They have few substitutes and are largely complements to the production of private-sector goods and services. Ashauer (1990) found that infrastructure investment is an important source of productivity growth and should be included as an input in the production function. As with R&D spending, the full impact of government infrastructure investment may extend well beyond the direct benefits of the projects, because improvements in the economy's infrastructure generally boost the productivity of private investments.

4.11. Summary

Long-term sustainable growth is determined by the rate of expansion of real potential GDP. Expansion of the supply of factors of production (inputs) and improvements in technology are the sources of growth. The factors of production include human capital, ICT and non-ICT capital, public capital, labor, and natural resources. Data for the sources of growth are available from the OECD and the Conference Board. Exhibit 11-12 provides data from the Conference Board on the sources of output growth for various countries. These estimates are based on the growth accounting formula.[19]

[19]A standard growth accounting model (expanded version of Equation 11-4) is used to compute the contribution of each input to aggregate output (GDP) growth. The inputs include both the quantity and quality of labor and ICT and non-ICT capital. Each input is weighted by its share in national income, and TFP captures all sources of growth that are left unexplained by the labor and capital inputs.

EXHIBIT 11-12 Sources of Output Growth

	Contribution from:					
	Labor Quantity (%)	Labor Quality (%)	Non-ICT Capital (%)	ICT Capital (%)	TFP (%)	Growth in GDP (%)
Germany						
1995–2005	−0.2	0.1	0.3	0.2	0.9	1.3
2005–2009	−0.6	0.1	0.5	0.3	0.1	0.4
Ireland						
1995–2005	2.0	0.3	2.6	0.7	1.7	7.3
2005–2009	−0.2	0.1	1.8	0.4	−2.1	0.0
United States						
1995–2005	0.6	0.3	0.7	0.8	0.9	3.3
2005–2009	0.1	0.1	0.5	0.5	−0.5	0.7
Japan						
1995–2005	−0.6	0.4	0.6	0.3	0.4	1.1
2005–2009	−0.6	0.1	0.4	0.2	−0.6	−0.5
South Korea						
1995–2005	−0.5	0.8	1.1	0.5	2.4	4.3
2005–2009	−0.7	0.0	0.8	0.2	2.0	2.3
China						
1995–2005	0.5	0.2	4.5	1.1	1.5	7.8
2005–2008	0.3	0.2	5.5	1.3	4.2	11.5
India						
1995–2005	1.0	0.2	2.7	0.5	1.9	6.3
2005–2008	1.1	0.1	3.7	0.9	2.4	8.2
Brazil						
1995–2005	0.8	0.1	1.1	0.7	−0.3	2.4
2005–2008	0.8	0.2	1.9	2.5	−0.5	4.9
Mexico						
1995–2005	1.2	0.2	1.4	0.4	0.4	3.6
2005–2008	1.1	0.1	1.3	0.2	−0.1	2.6

Source: Conference Board Total Economy Database.

EXAMPLE 11-8 The Irish Economy

As shown in Exhibit 11-1, economic growth in Ireland since 1970 has been significantly higher than that experienced in the major European economies of Germany, France, and the United Kingdom. In 1970, the per capita GDP of Ireland, at $9,869, was 45.2 percent below the per capita income of the United Kingdom. In 2010, per capita GDP in Ireland, at $36,433, was only 2.5 percent below the United Kingdom's $37,371 per capita GDP. Like most of the global economy, Ireland fell into a deep recession in 2009, with GDP contracting by over 7 percent. To understand the factors driving the Irish economy and the prospects for future equity returns, use the data in Exhibits 11-11 and 11-12 and the following population data to address these questions:

1. Using the growth accounting framework data, evaluate the sources of growth for the Irish economy from 1995 to 2009.
2. What is likely to happen to the potential rate of growth for Ireland? What are the prospects for equity returns?

	2000	2010	Annual Growth Rate
Population (millions)	3.8	4.5	1.71%
Net immigration total (2000–2010)		375,587	
Net immigration total (2009–2010)		−20,000	
Population less immigrants (millions)	3.8	4.1	0.8%

Solution to 1: The sources of growth for an economy include labor quantity, labor quality, non-ICT capital, ICT capital, and TFP. The growth accounting data in Exhibit 11-12 indicate that economic growth in Ireland from 1995 to 2009 is explained by the following factors:

Input	Contribution: 1995–2005	Contribution: 2005–2009
Labor	**2.3%**	**−0.1%**
Labor quantity	2.0%	−0.2%
Labor quality	0.3%	0.1%
Capital/Investment	**3.3%**	**2.2%**
Non-ICT capital	2.6%	1.8%
ICT capital	0.7%	0.4%
TFP	**1.7%**	**−2.1%**
Total: GDP growth	**7.3%**	**0.0%**

In sum, the main driver of growth for the Irish economy since 1995 has been capital spending. It accounted for over 45 percent of growth in 1995–2005 and has been the only factor contributing to growth in the Irish economy since 2005, offsetting the negative contribution from labor and TFP. Another way to look at growth in Ireland for the period 2005–2009 is that all the growth is through capital deepening. As shown in Exhibit 11-11, capital deepening added 2.9 percent to growth and by offsetting the decline in TFP caused an increase in labor productivity of 0.8 percent.

Solution to 2: Looking forward, prospects for the economy are not as favorable as in the past. To estimate the growth rate in potential GDP, we repeat Equation 11-5, given by:

$$\begin{aligned}\text{Growth rate of potential GDP} = {} & \text{Long-term growth rate of labor force} \\ & + \text{Long-term growth rate in labor productivity}\end{aligned}$$

To use this equation, we need to project the growth rate in the labor input and labor productivity. The total hours worked data in Exhibit 11-11 are one potential source to use to estimate the growth rate of the labor input. Exhibit 11-11 shows the labor input declining by 0.8 percent between 2005 and 2009. The problem here is that the decline in hours worked is overstated because of the negative impact of the recession on hours worked. As an alternative, the labor input should grow at the same rate as the population plus the net change due to immigration. The population data for Ireland (given after Question 2) show that over half of the population growth between 2000 and 2010 was due to immigration. Since 2009, however, outward migration has replaced inward migration, reducing the growth rate in the labor input. Thus, if the 2000–2010 influx of immigrants is reversed over the next decade, a reasonable, perhaps somewhat conservative, estimate for labor force growth is zero. We also assume:

- There is no increase in labor productivity coming from capital deepening as investment slows (resulting in essentially no growth in net investment and the physical capital stock).
- TFP growth reverts to its average growth rate of 1.7 percent in the 1995–2005 time period (see Exhibit 11-11).
- Labor productivity grows at the same rate as TFP.

Thus, growth in potential GDP is 0.0% + 1.7% = 1.7%.

In summary, despite the projected rebound in TFP growth, overall potential growth in Ireland is likely to decline as labor input growth and capital deepening no longer contribute to overall growth. As discussed in Section 3 of the chapter, slower growth in potential GDP will limit potential earnings growth and equity price appreciation.

EXAMPLE 11-9 Investment Outlook for China and India

The investment policy committee at Global Invest Inc. is interested in increasing the firm's exposure to either India or China because of their rapid rates of economic growth. Economic growth in China has been close to 10 percent over the past few years, and India has grown over 7 percent. You are asked by the committee to do the following:

1. Determine the sources of growth for the two economies and review the data on productivity and investment using information from Exhibits 11-5, 11-9, 11-10, 11-11, and 11-12. Which of the two countries looks more attractive based on the sources of growth?
2. Estimate the long-term sustainable earnings growth rate using data from 1995 to 2008.
3. Make an investment recommendation.

Solution to 1: The sources of economic growth include size of labor force, quality of labor force (human capital), ICT and non-ICT capital, natural resources, and technology. Looking at the sources of growth in Exhibit 11-12, we get the following:

Input	Percent Contribution: 1995–2005	Percent Contribution: 2005–2008
India		
Labor quantity	1.0	1.1
Labor quality	0.2	0.1
Non-ICT capital	2.7	3.7
ICT capital	0.5	0.9
TFP	1.9	2.4
Total: GDP growth	6.3	8.2
China		
Labor quantity	0.5	0.3
Labor quality	0.2	0.2
Non-ICT capital	4.5	5.5
ICT capital	1.1	1.3
TFP	1.5	4.2
Total: GDP growth	7.8	11.5

- The contribution of the labor quantity input is more important to growth in India than in China. Labor quantity contributed 1 percent to India's GDP growth over 1995–2005 and 1.1 percent over 2005–2008. The equivalent numbers for China are 0.5 percent and 0.3 percent, respectively. Looking ahead, labor is likely to be a major factor adding to India's growth. The population of India (Exhibit 11-5) is growing at a faster rate than that of China. The annual growth rate in population from 2001 to 2010 was 1.52 percent in India versus 0.97 percent in China. Also, hours worked in India (Exhibit 11-11) are growing at a faster rate than in China. Therefore, the workforce and labor quantity input should grow faster in India. The edge here goes to India.
- The contribution to GDP made by the quality of the labor force is essentially identical in the two countries (0.2 percent in China versus 0.2 percent in India between 1995 and 2005 and 0.2 percent in China and 0.1 percent in India between 2005 and 2008). This factor is a tie.
- The contribution of non-ICT capital investment is significantly higher in China (4.5 percent in China versus 2.7 percent in India between 1995 and 2005 and 5.5 percent in China and 3.7 percent in India between 2005 and 2008). The edge goes to China.
- The contribution of ICT capital investment is significantly higher in China (1.1 percent in China versus 0.5 percent in India between 1995 and 2005 and 1.3 percent in China and 0.9 percent in India between 2005 and 2008). The edge goes to China.
- Both countries spend a high percentage of GDP on capital investment (Exhibit 11-9). In 2010, investment spending as a percentage of GDP was 48.2 percent in China and 36.8 percent in India. The Chinese share is higher, and this provides China with an edge unless diminishing marginal returns to capital deepening become an issue. However, this is not likely for a while given the low level of capital per worker in China. China and India still have a long way to go to converge with the developed economies. The advantage goes to China.
- The contribution of technological progress is measured by TFP. Comparing the two countries, TPF growth was higher in India over the period 1995–2005 (1.9 percent in India versus 1.5 percent in China). For the period 2005–2009, however, TFP growth was significantly higher in China (4.2 percent versus 2.4 percent). In addition, expenditures on R&D for 2009 (Exhibit 11-10) as a percentage of GDP were higher in China (1.7 percent in China and 0.8 percent in India). The edge here goes to China.
- Finally, growth in overall labor productivity (Exhibit 11-11) is considerably higher in China than India (10.3 percent in China versus 6.0 percent in India between 2005 and 2008). This is due to a greater increase in the capital-to-labor ratio in China (because of the high rate of investment, the physical capital stock is growing faster than the labor input) and due to faster technological progress in China. The edge here goes to China.

In sum, based on the sources of growth, China appears to be better positioned for growth in the future.

Solution to 2: Estimates of potential GDP using the inputs from Exhibit 11-11 for China and India are:

Growth rate in potential GDP = Long-term growth rate of labor force (equals growth in hours worked in Exhibit 11-11) + Long-term growth rate in labor productivity

China (using 1995–2008)*

$$\text{Growth in potential} = 1.1\% + 7.5\% = 8.6\%$$

India (using 1995–2008)

$$\text{Growth in potential} = 2.1\% + 4.6\% = 6.7\%$$

Solution to 3: Growth prospects in both countries are very attractive. However, China's growth potential is higher because of its greater level of capital spending and the greater contribution of technological progress toward growth. Long-term earnings growth is closely tied to the growth rate in potential GDP. Therefore, based on the previous calculations, earnings in China would be projected to grow at an annual rate of 8.6 percent, compared with 6.7 percent in India. Over the next decade, ignoring current valuation, the Chinese equity market would be projected to outperform the Indian market as its higher rate of sustainable growth translates into a higher rate of appreciation in equity values.†

*Calculated as geometric mean growth rates using data for the 1995–2005 and 2005–2008 subperiods.

†It bears repeating that the global economy is evolving rapidly and past trends may or may not be sustained. This is especially true of China and India. To provide concrete answers that do not require the reader to bring in additional information, our exercise solutions must assume past patterns are indicative of the future.

5. THEORIES OF GROWTH

The factors that drive long-term economic growth and determine the rate of sustainable growth in an economy are the subject of much debate among economists. The academic growth literature includes three main paradigms with respect to per capita growth in an economy—the classical, neoclassical, and endogenous growth models. Per capita economic growth under the classical model is only temporary because an exploding population with limited resources brings growth to an end. In the neoclassical model, long-run per capita growth depends solely on exogenous technological progress. The final model of growth attempts to explain technology within the model itself—thus the term endogenous growth.

5.1. Classical Model

Classical growth theory was developed by Thomas Malthus in his 1798 publication *Essay on the Principle of Population.* Commonly referred to as the Malthusian theory, it is focused on the impact of a growing population in a world with limited resources. The concerns of resource depletion and overpopulation are central themes within the Malthusian perspective on growth. The production function in the classical model is relatively simple and consists of a labor input with land as a fixed factor. The key assumption underlying the classical model is that population growth accelerates when the level of per capita income rises above the subsistence income, which is the minimum income needed to maintain life. This means that technological progress and land expansion, which increase labor productivity, translate into higher population growth. But because the labor input faces diminishing marginal returns, the additional output produced by the growing workforce eventually declines to zero. Ultimately, the population grows so much that labor productivity falls and per capita income returns back to the subsistence level.

The classical model predicts that in the long run, the adoption of new technology results in a larger but not richer population. Thus, the standard of living is constant over time even with technological progress, and there is no growth in per capita output. As a result of this gloomy forecast, it is thought, economics was labeled the "dismal science."

The prediction from the Malthusian model failed for two reasons:

1. The link between per capita income and population broke down. In fact, as the growth of per capita income increased, population growth slowed rather than accelerating as predicted by the classical growth model.
2. Growth in per capita income has been possible because technological progress has been rapid enough to more than offset the impact of diminishing marginal returns.

Because the classical model's pessimistic prediction never materialized, economists changed the focus of the analysis away from labor to capital and to the neoclassical model.

5.2. Neoclassical Model

Robert Solow devised the mainstream neoclassical theory of growth in the 1950s.[20] The heart of this theory is the Cobb–Douglas production function discussed in Section 4.1. As before, the potential output of the economy is given by:

$$Y = AF(K, L) = AK^{\alpha}L^{1-\alpha}$$

where K is the stock of capital, L is the labor input, and A is total factor productivity.[21] In the neoclassical model, both capital and labor are variable inputs each subject to diminishing marginal productivity.

[20] Solow (1957).

[21] Our exposition of the neoclassical model with technological progress reflected in total factor productivity corresponds to what is known as "Hicks neutral" technical change. The neoclassical model is usually presented with "Harrod neutral" or "labor augmenting" technical change. In that formulation, the production function is given by $Y = F(K, BL)$, where B represents technological change and (BL) is interpreted as the effective labor supply. In general, this is not equivalent to our formulation using TFP. However, they are equivalent if, as we assume here, the function F( ) has the Cobb–Douglas form. To see this, note that $[K^{\alpha}(BL)^{1-\alpha}] = [B^{1-\alpha}(K^{\alpha}L^{1-\alpha})] = [A(K^{\alpha}L^{1-\alpha})]$, where $A \equiv B^{1-\alpha}$ is total factor productivity.

The objective of the neoclassical growth model is to determine the long-run growth rate of output per capita and relate it to (1) the savings/investment rate, (2) the rate of technological change, and (3) population growth.

5.2.1. Balanced or Steady State Rate of Growth

As with most economic models, the neoclassical growth model attempts to find the equilibrium position toward which the economy will move. In the case of the Solow model, this equilibrium is the balanced or **steady state rate of growth** that occurs when the output-to-capital ratio is constant. Growth is balanced in the sense that capital per worker and output per worker grow at the same rate.

We begin the analysis by using the per capita version of the Cobb–Douglas production function given earlier in Equation 11-3:

$$y = Y/L = Ak^{\alpha}$$

where $k = K/L$. Using the Cobb–Douglas definitions, the rates of change of capital per worker and output per worker are given by:[22]

$$\Delta k/k = \Delta K/K - \Delta L/L$$

and:

$$\Delta y/y = \Delta Y/Y - \Delta L/L$$

From the production function, the growth rate of output per worker is also equal to:

$$\Delta y/y = \Delta A/A + \alpha \Delta k/k \tag{11-6}$$

The physical capital stock in an economy will increase because of gross investment (I) and decline because of depreciation. In a closed economy, investment must be funded by domestic saving. Letting s be the fraction of income (Y) that is saved, gross investment is given by $I = sY$. Assuming the physical capital stock depreciates at a constant rate, δ, the change in the physical capital stock is given by:

$$\Delta K = sY - \delta K$$

Subtracting labor supply growth, $\Delta L/L \equiv n$, and rearranging gives:

$$\Delta k/k = sY/K - \delta - n \tag{11-7}$$

In the steady state, the growth rate of capital per worker is equal to the growth rate of output per worker. Thus,

$$\Delta k/k = \Delta y/y = \Delta A/A + \alpha \Delta k/k$$

[22]Strictly speaking, these and other rate of change equations are exact only for changes over arbitrarily short periods (continuous time).

from which we get:

$$\Delta y/y = \Delta k/k = (\Delta A/A)/(1 - \alpha)$$

Letting θ denote the growth rate of TFP (i.e., $\Delta A/A$), we see that the equilibrium sustainable growth rate of output per capita (= Growth rate of capital per worker) is a constant that depends only on the growth rate of TFP (θ) and the elasticity of output with respect to capital (α). Adding back the growth rate of labor (n) gives the sustainable growth rate of output.

$$\begin{aligned} \text{Growth rate of output per capita} &= \frac{\theta}{1 - \alpha} \\ \text{Growth rate of output} &= \frac{\theta}{1 - \alpha} + n \end{aligned} \tag{11-8}$$

This is the key result of the neoclassical model. Note that $[\theta/(1 - \alpha)]$ is the steady state growth rate of labor productivity, so Equation 11-8 is consistent with the labor productivity growth accounting equation discussed in Section 4.3.

Substituting $[\theta/(1 - \alpha)]$ into the left-hand side of Equation 11-7 and rearranging gives the equilibrium output-to-capital ratio, denoted by the constant Ψ.

$$\frac{Y}{K} = \left(\frac{1}{s}\right)\left[\left(\frac{\theta}{1 - \alpha}\right) + \delta + n\right] \equiv \Psi \tag{11-9}$$

In the steady state, the output-to-capital ratio is constant and the capital-to-labor ratio (k) and output per worker (y) grow at the same rate, given by $[\theta/(1 - \alpha)]$. On the steady state growth path, the marginal product of capital is also constant and, given the Cobb–Douglas production function, is equal to $\alpha(Y/K)$. The marginal product of capital is also equal to the real interest rate in the economy. Note that even though the capital-to-labor ratio (k) is rising at rate $[\theta/(1 - \alpha)]$ in the steady state, the increase in the capital-to-labor ratio (k) has no impact on the marginal product of capital, which is not changing. *Capital deepening is occurring, but it has no effect on the growth rate of the economy or on the marginal product of capital once the steady state is reached.*

EXAMPLE 11-10 Steady State Rate of Growth for China, Japan, and Ireland

Earlier examples generated estimates of potential growth for China (11.5 percent), Japan (0.8 percent), and Ireland (1.7 percent). Given the following data,

1. Calculate the steady state growth rates from the neoclassical model for China, Japan, and Ireland.

2. Compare the steady state growth rates to the growth rates in potential GDP estimated in Examples 11-7 to 11-9 and explain the results.

	Labor Cost in Total Factor Cost (%)	TFP Growth (%)	Labor Force Growth (%)
China	46.5	2.5	1.2
Japan	57.3	0.2	0.0
Ireland	56.7	0.8	0.0

Sources: Conference Board Total Economy Database; labor cost and TFP growth are based on 1995–2009 data for Japan and Ireland and 1995–2008 data for China. Labor force growth estimates are from earlier examples.

Solution to 1: Using Equation 11-8, the steady state growth rate in the neoclassical model is given by:

$$\begin{aligned}\Delta Y/Y &= (\theta)/(1-\alpha) + n \\ &= \text{Growth rate of TFP scaled by labor factor share} \\ &\quad + \text{Growth rate in the labor force}\end{aligned}$$

Using the preceding equation and data, steady state growth rates for the three countries are estimated as follows:

China: The labor share of output $(1 - \alpha)$ is given by the average of the labor cost as a percentage of total factor cost, which is equal to 0.465 for China. The growth rate in the labor force is 1.2 percent, and the growth rate of TFP is 2.5 percent.

$$\text{Steady state growth rate} = 2.5\%/0.465 + 1.2\% = 6.58\%$$

Japan: The labor share of output $(1 - \alpha)$ for Japan is 0.573. The growth rate in the labor force is 0.0 percent, and TFP growth is 0.2 percent.

$$\text{Steady state growth rate} = 0.2\%/0.573 + 0.0\% = 0.35\%$$

Ireland: The labor share of output $(1 - \alpha)$ is 0.567 percent for Ireland. The growth rate in the labor force is 0.0 percent, and TFP growth is 0.8 percent.

$$\text{Steady state growth rate} = 0.8\%/0.567 + 0.0\% = 1.41\%$$

Solution to 2: The growth rate in potential GDP for China (8.6 percent, estimated in Example 11-9) is significantly above the estimated 6.58 percent steady state growth rate. The reason for this is that the economy of China is still in the process of converging to the higher income levels of the United States and the major economies in Europe. The

physical capital stock is below the steady state, and capital deepening is a significant factor increasing productivity growth (see Exhibit 11-11) and the growth in potential GDP.

This is not the case for Japan and Ireland. Both countries are operating at essentially the steady state. The estimated growth rate in potential GDP for Japan (0.8 percent, from Example 11-7) is only slightly above its 0.35 percent steady state growth rate. Likewise, the estimated growth rate in potential GDP for Ireland (1.7 percent, from Example 11-8) is effectively equal to its estimated steady state growth rate of 1.4 percent. Operating close to the steady state means that capital investment in these countries, which results in an increasing capital-to-labor ratio, has no significant effect on the growth rate of the economy. Only changes in the growth rates of TFP and labor and in the labor share of output have an impact on potential GDP growth.

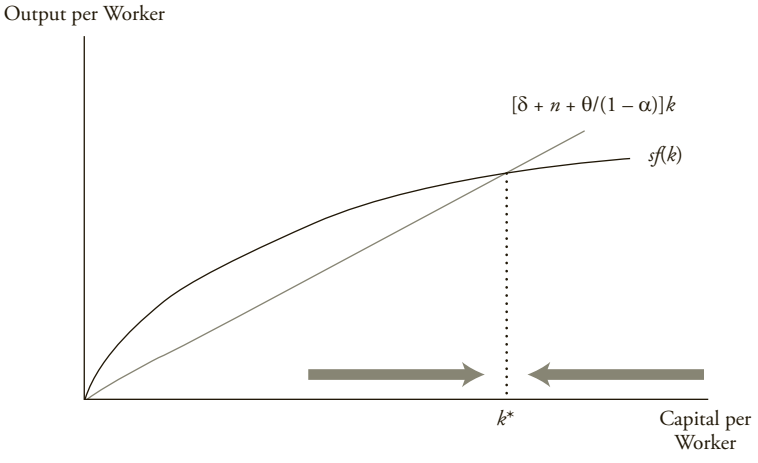
An intuitive way to understand the steady state equilibrium given in Equation 11-9 is to transform it into a savings/investment equation:

$$sy = \left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right]k$$

Steady state equilibrium occurs at the output-to-capital ratio where the savings and actual gross investment per worker generated in the economy (sy) are just sufficient to (1) provide capital for new workers entering the workforce at rate n, (2) replace plant and equipment wearing out at rate δ, and (3) deepen the physical capital stock at the rate $[\theta/(1-\alpha)]$ required to keep the marginal product of capital equal to the rental price of capital.

Exhibit 11-13 shows the steady state equilibrium graphically. The straight line in the exhibit indicates the amount of investment required to keep the physical capital stock

EXHIBIT 11-13 Steady State in the Neoclassical Model

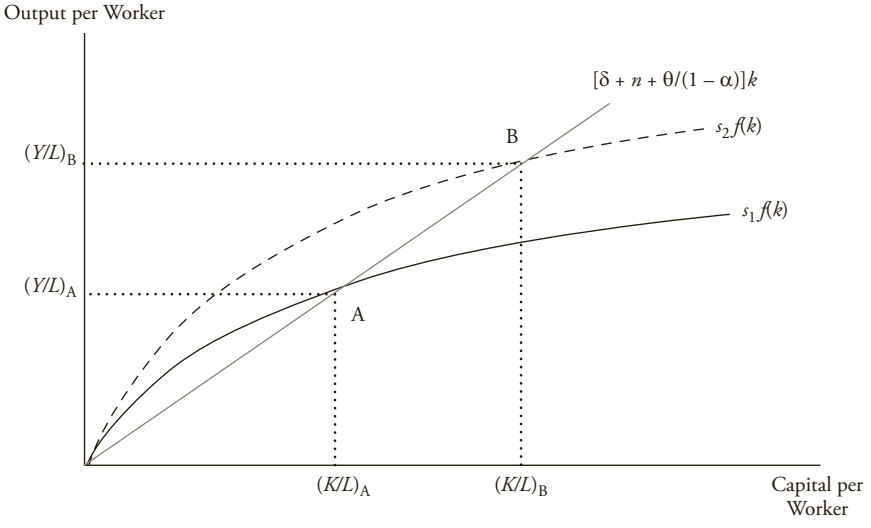


growing at the required rate. Because the horizontal axis is capital per worker, the slope of the line is given by $[\delta + n + \theta/(1 - \alpha)]$. The curved line shows the amount of actual investment per worker and is determined by the product of the saving rate and the production function. It is curved because of diminishing marginal returns to the capital input in the production function. The intersection of the required investment and actual investment lines determines the steady state. Note that *this exhibit is a snapshot at a point in time*. Over time, the capital-to-labor ratio rises at rate $[\theta/(1 - \alpha)]$ as the actual saving/investment curve $[sf(k)]$ shifts upward because of TFP growth, and *the equilibrium moves upward and to the right along the straight line.*

The impact of the various parameters in the model on the steady state can also be seen in the exhibit. At any point in time when the economy is on its steady state growth path, the exogenous factors—labor supply and TFP—are fixed. We would like to know what effect each of the parameters in the model has on the steady state capital-to-labor ratio and therefore on output per worker. For example, if there are two economies that differ only with respect to one parameter, what does that imply about their per capita incomes? All else the same, we can say the following regarding the impact of the parameters:

- Saving rate (s): An increase in the saving rate implies a higher capital-to-labor ratio (k) and higher output per worker (y) because a higher saving rate generates more saving/investment at every level of output. In Exhibit 11-14, the saving/investment curve $[sf(k)]$ shifts upward from an initial steady state equilibrium at point A to a new equilibrium at point B. At the new equilibrium point, it intersects the required investment line $[\delta + n + \theta/(1 - \alpha)]$ at higher capital-to-labor and output per worker ratios. Note that although the higher saving rate increases both k and y, it has no impact on the steady state growth rates of output per capita or output (Equation 11-8).
- Labor force growth (n): An increase in the labor force growth rate reduces the equilibrium capital-to-labor ratio because a corresponding increase in the steady state growth rate of

EXHIBIT 11-14 Impact on the Steady State: Increase in the Saving Rate



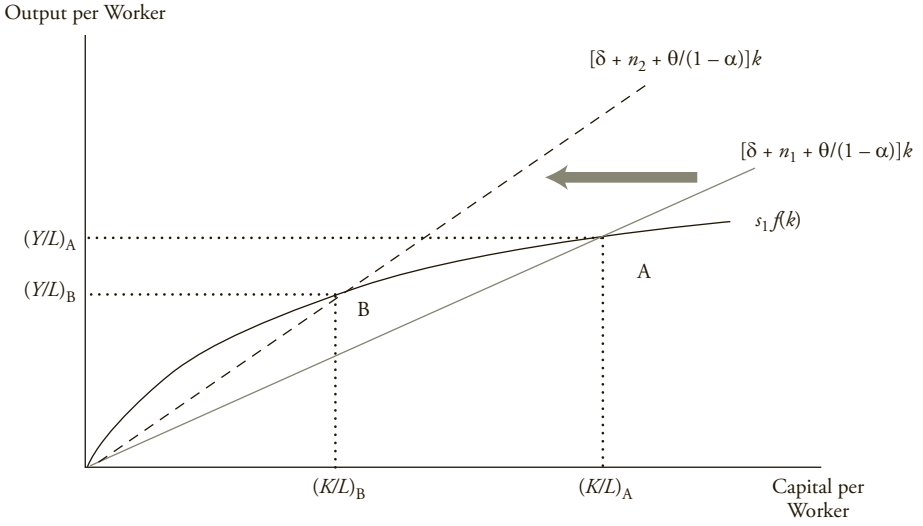
capital is required. Given the gross saving/investment rate, this can only be achieved at a lower capital-to-labor ratio. Output per worker is correspondingly lower as well. In Exhibit 11-15, the higher population growth rate increases the slope of the required investment line. This shifts the steady state equilibrium from point A to point B, where it intersects the supply of saving/investment curve at lower capital-to-labor and output per worker ratios.

- Depreciation rate (δ): An increase in the depreciation rate reduces the equilibrium capital-to-labor and output per worker ratios because a given rate of gross saving generates less net capital accumulation. Graphically, it increases the slope of the required investment line and affects the steady state equilibrium in the same way as labor force growth (Exhibit 11-15).
- Growth in TFP (θ): An increase in the growth rate of TFP reduces the steady state capital-to-labor ratio and output per worker for given levels of labor input and TFP. This result must be interpreted with care. Raising the growth rate of TFP means that output per worker will grow faster in the future (Equation 11-8), but at a given point in time, a given supply of labor, and a given *level* of TFP, output per worker is lower than it would be with a slower TFP growth rate. In effect, the economy is on a steeper trajectory off a lower base of output per worker. Graphically, it is identical to Exhibit 11-15 in that faster TFP growth steepens the required investment line (increases the slope), which intersects with the available saving/investment curve at lower capital-to-labor and investment per worker ratios.

In sum, such factors as the saving rate, the growth rate of the labor force, and the depreciation rate change the *level* of output per worker but do not permanently change the *growth rate* of output per worker. A permanent increase in the growth rate in output per worker can occur only if there is a change in the growth rate of TFP.

So far we have focused on the steady state growth path. What happens if the economy has not yet reached the steady state? During the transition to the steady state growth path, the economy can experience either faster or slower growth relative to the steady state. Using Equations 11-6, 11-7, and 11-9, we can write the growth rates of output per capita and the capital-to-labor ratio as, respectively,

EXHIBIT 11-15 Impact on the Steady State: Increase in Labor Force Growth



$$\frac{\Delta y}{y} = \left(\frac{\theta}{1-\alpha}\right) + \alpha s\left(\frac{Y}{K} - \psi\right) = \left(\frac{\theta}{1-\alpha}\right) + \alpha s(y/k - \psi) \tag{11-10}$$

and

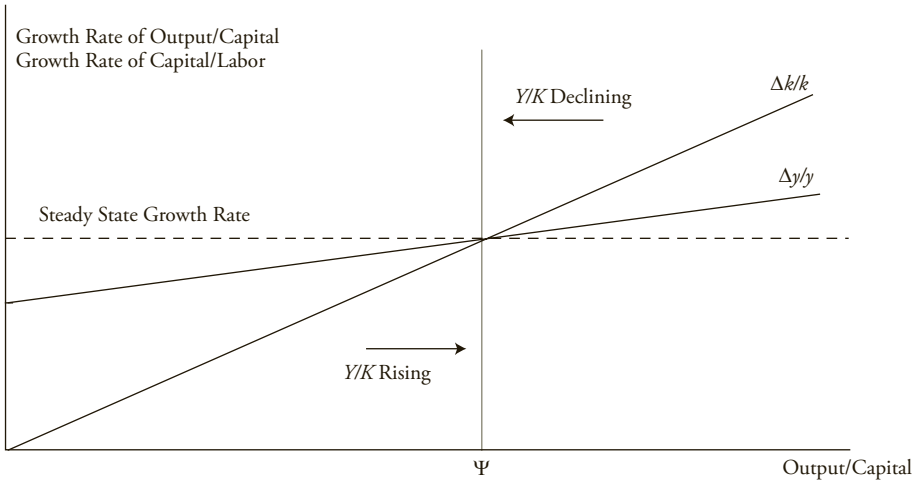
$$\frac{\Delta k}{k} = \left(\frac{\theta}{1-\alpha}\right) + s\left(\frac{Y}{K} - \psi\right) = \left(\frac{\theta}{1-\alpha}\right) + s(y/k - \psi), \tag{11-11}$$

where the second equality in each line follows from the definitions of y and k, which imply $(Y/K) = y/k$. These relationships are shown in Exhibit 11-16.

If the output-to-capital ratio is above its equilibrium level (ψ), the second term in Equations 11-10 and 11-11 is positive and the growth rates of output per capita and the capital-to-labor ratio are above the steady state rate $[\theta/(1 - \alpha)]$. This corresponds to a situation in which actual saving/investment exceeds required investment and above-trend growth in per capita output is driven by an above-trend rate of capital deepening. This situation usually reflects a relatively low capital-to-labor ratio but could, at least in principle, arise from high TFP. Because $\alpha < 1$, capital is growing faster than output and the output-to-capital ratio is falling. Over time, the growth rates of both output per capita and the capital-to-labor ratio decline to the steady state rate.

Of course, the converse is true if the output-to-capital ratio is below its steady state level. Actual investment is insufficient to sustain the trend rate of growth in the capital-to-labor ratio, and both output per capita and the capital-to-labor ratio grow more slowly. This situation usually corresponds to a relatively high and unsustainable capital-to-labor ratio, but could reflect relatively low TFP and hence relatively low output. Over time, output grows faster than capital, the output-to-capital ratio rises, and growth converges to the trend rate.

EXHIBIT 11-16 Dynamics in the Neoclassical Model



5.2.2. Implications of the Neoclassical Model

There are four major groups of conclusions from the neoclassical model:

1. Capital accumulation:
   - Capital accumulation affects the level of output but not the growth rate in the long run.
   - Regardless of its initial capital-to-labor ratio or initial level of productivity, a growing economy will move to a point of steady state growth.
   - In a steady state, the growth rate of output equals the rate of labor force growth plus the rate of growth in TFP scaled by labor's share of income $[n + \theta/(1 - \alpha)]$.[23] The growth rate of output does not depend on the accumulation of capital or the rate of business investment.
2. Capital deepening versus technology:
   - Rapid growth that is above the steady state rate of growth occurs when countries first begin to accumulate capital; but growth will slow as the process of accumulation continues (see Exhibit 11-16).
   - Long-term sustainable growth cannot rely solely on capital deepening investment—that is, on increasing the stock of capital relative to labor. If the capital-to-labor ratio grows too rapidly (i.e., faster than labor productivity), capital becomes less productive, resulting in slower rather than faster growth.
   - More generally, increasing the supply of some input(s) too rapidly relative to other inputs will lead to diminishing marginal returns and cannot be the basis for sustainable growth.
   - In the absence of improvements in TFP, the growth of labor productivity and per capita output would eventually slow.
   - Because of diminishing marginal returns to capital, the only way to sustain growth in potential GDP per capita is through technological change or growth in total factor productivity. This results in an upward shift in the production function—the economy produces more goods and services for any given mix of labor and capital inputs.
3. Convergence:
   - Given the relative scarcity and hence high marginal productivity of capital and potentially higher saving rates in developing countries, the growth rates of developing countries should exceed those of developed countries.
   - As a result, there should be a convergence of per capita incomes between developed and developing countries over time.
4. Effect of savings on growth:
   - The initial impact of a higher saving rate is to temporarily raise the rate of growth in the economy.[24] In response to the higher saving rate, growth exceeds the steady state

[23]Readers who are familiar with the labor-augmenting technical change formulation of the neoclassical model should note that in that formulation the rate of labor-augmenting technical change is also the growth rate of labor productivity. In our formulation, the growth rate of labor productivity is $[\theta/(1 - \alpha)]$. So both formulations imply that long-run growth equals the growth rate of the labor supply (n) plus a constant growth rate of labor productivity.

[24]Mathematically, this can be seen as follows: Equation 11-9 indicates that an increase in the saving rate (s) reduces the steady state output-to-capital ratio (ψ). This makes the last term in Equations 11-10 and 11-11 positive, raising the growth rates of output per capita (y) and the capital-to-labor ratio (k) above the steady state rate.

growth rate during a transition period. However, the economy returns to the balanced growth path after the transition period.
- During the transition period, the economy moves to a higher level of per capita output and productivity.
- Once an economy achieves steady state growth, the growth rate does not depend on the percentage of income saved or invested. Higher savings cannot permanently raise the growth rate of output.
- However, countries with higher saving rates will have a higher level of per capita output, a higher capital-to-labor ratio, and a higher level of labor productivity.

EXAMPLE 11-11 Comparative Statics and Transitional Growth in the Neoclassical Model

Beginning in steady state equilibrium, an economy's saving rate suddenly increases from 20 percent of income to 30 percent of income. Other key parameters describing the economy are:

Growth rate of TFP (θ)	$= 0.02$
Income share of capital (α)	$= 0.35$
Depreciation rate (δ)	$= 0.10$
Labor force growth rate (n)	$= 0.01$

The following table shows the output-to-capital ratio that will prevail in this economy at various points in time after the increase in the saving rate.

Years after Saving Rate Increase	Output-to-Capital Ratio
5	0.5947
10	0.5415
25	0.4857
50	0.4708
100	0.4693
New steady state	??

By rearranging the Cobb–Douglas production function (Equation 11-3), the proportional impact of the saving rate change on the capital-to-labor ratio can be expressed in terms of the proportional impact on the output-to-capital ratio. The proportional impact on per capita income can then be determined from the production function

(Equation 11-3). Labeling the paths with and without the change in saving rate as *new* and *old*, respectively, at each date we have:*

$$\frac{k_{new}}{k_{old}} = \left[\frac{(Y/K)_{new}}{(Y/K)_{old}}\right]^{\frac{1}{\alpha-1}}$$

and:

$$\frac{y_{new}}{y_{old}} = \left(\frac{k_{new}}{k_{old}}\right)^{\alpha}$$

1. Using Equations 11-8 and 11-9, calculate the steady state growth rate of per capita income and the steady state output-to-capital ratio both before and after the change in the saving rate. What happens to the capital-to-labor ratio and output per capita?
2. Use the output-to-capital ratios given in the table above along with Equation 11-10 and your answers to Question 1 to determine the growth rate of per capita income that will prevail immediately following the change in the saving rate and at each of the indicated times after the change. Explain the pattern of growth rates.
3. Using the output-to-capital ratios given in the table, calculate the proportional impact of the increased saving rate on the capital-to-labor ratio and on per capita income over time. With respect to these variables, how will the new steady state compare with the old steady state?

Solution to 1: From Equation 11-8, the steady state growth rate of per capita income, both before and after the increase in the saving rate, is $\Delta y/y = \theta/(1 - \alpha) = 0.02/(1 - 0.35) = 0.0308$, or 3.08 percent. From Equation 11-9, the steady state output-to-capital ratio is:

$$\frac{Y}{K} = \left(\frac{1}{s}\right)\left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right] \equiv \Psi$$

Using the parameter values given earlier, $[\theta/(1 - \alpha) + \delta + n] = (0.0308 + 0.10 + 0.01) = 0.1408$, so the steady state output-to-capital ratio is $(0.1408/0.2) = 0.7040$ with the initial 20 percent saving rate and $(0.1408/0.30) = 0.4693$ with the new 30 percent saving rate. As shown in Exhibit 11-14, both the capital-to-labor ratio and output per worker are at higher *levels* in the new steady state. But once the new steady state is achieved, they do not grow any faster than they did in the steady state with the lower saving rate.

Solution to 2: According to Equation 11-10, the growth rate of per capita income is given by:

$$\frac{\Delta y}{y} = \left(\frac{\theta}{1-\alpha}\right) + \alpha s(y/k - \psi)$$

Immediately following the increase in the saving rate, the relevant value of ψ becomes the new steady state output-to-capital ratio (0.4693). The actual output-to-

capital ratio does not change immediately, so y/k is initially still 0.7040. Plugging these values into the growth equation gives the growth rate of per capita income:

$$\Delta y/y = 0.0308 + (0.35)(0.30)(0.7040 - 0.4693) = 0.0554, \text{ or } 5.54\%$$

Similar calculations using the output-to-capital ratios in the preceding table give the following:

Years after Saving Rate Increase	Output-to-Capital Ratio	Growth Rate of Per Capita Income (%)
0	0.7040	5.54
5	0.5947	4.39
10	0.5415	3.84
25	0.4857	3.25
50	0.4708	3.09
100	0.4693	3.08
New steady state	0.4693	3.08

The growth rate jumps from the steady state rate of 3.08 percent to 5.54 percent when the saving rate increases, because the increase in saving/investment results in more rapid capital accumulation. Over time, the growth rate slows because the marginal productivity of capital declines as the capital-to-labor ratio increases. In addition, as the capital-to-labor ratio increases and the output-to-capital ratio declines, a greater portion of savings is required to maintain the capital-to-labor ratio, leaving a smaller portion for continued capital deepening. Roughly two-thirds of the growth acceleration has dissipated after 10 years.

Solution to 3: Using the output-to-capital ratio that will prevail five years after the saving rate increase, the proportional impact on the capital-to-labor ratio and on per capita income will be:

$$\frac{k_{new}}{k_{old}} = \left[\frac{(Y/K)_{new}}{(Y/K)_{old}}\right]^{\frac{1}{\alpha-1}} = \left(\frac{0.5947}{0.7040}\right)^{\frac{-1}{0.65}} = 1.2964$$

and:

$$\frac{y_{new}}{y_{old}} = \left(\frac{k_{new}}{k_{old}}\right)^{\alpha} = 1.2964^{0.35} = 1.0951$$

Thus, after five years, the capital-to-labor ratio will be 29.64 percent higher than it would have been without the increase in the saving rate, and per capita income will be 9.51 percent higher. Similar calculations for the other time periods give the following:

	Proportionate Increase (%) in:	
Years after Saving Rate Increase	Capital-to-Labor Ratio	Per Capita Income
0	0.00	0.00
5	29.64	9.51
10	49.74	15.18
25	77.01	22.12
50	85.71	24.19
100	86.68	24.42
New steady state	86.68	24.42

In the new steady state, the capital-to-labor ratio will be 86.68 percent higher at every point in time than it would have been in the old steady state. Per capita income will be 24.42 percent higher at every point in time. Both variables will be growing at the same rate (3.08 percent) as they would have been in the old steady state.

*Note that the output-to-capital ratio would have been constant on the original steady state path. Because of the impact of total factor productivity, the capital-to-labor ratio and output per capita are not constant even in steady state. In comparing paths for these variables, we isolate the impact of the saving rate change by canceling out the effect of TFP growth. Mathematically, we cancel out A in Equation 11-3 to get the equations shown here.

5.2.3. Extension of the Neoclassical Model

Solow (1957) used the growth accounting equation to determine the contributions of each factor to economic growth in the United States for the period 1909–1949. He reached the surprising conclusion that over 80 percent of the per capita growth in the United States was due to TFP. In another study, Denison (1985) examined U.S. growth for the period 1929–1982 using the Solow framework. His findings were similar to Solow's, with TFP explaining nearly 70 percent of U.S. growth. The problem with these findings is that the neoclassical model provides no explicit explanation of the economic determinants of technological progress or how TFP changes over time. Because technology is determined outside the model (i.e., exogenously), critics argue that the neoclassical model ignores the very factor driving growth in the economy. Technology is simply the residual or the part of growth that cannot be explained by other inputs, such as capital and labor. This lack of an explanation for technology led to growing dissatisfaction with the neoclassical model.

The other source of criticism of the neoclassical model is the prediction that the steady state rate of economic growth is unrelated to the rate of saving and investment. Long-run

growth of output in the Solow model depends only on the rates of growth of the labor force and technology. Higher rates of investment and savings have only a transitory impact on growth. Thus, an increase in investment as a share of GDP from 10 percent to 15 percent of GDP will have a positive impact on the near-term growth rate but will not have a permanent impact on the ultimately sustainable percentage growth rate. This conclusion makes many economists uncomfortable. Mankiw (1995) provided evidence rebutting this hypothesis and showed that saving rates and growth rates are positively correlated across countries. Finally, the neoclassical model predicts that in an economy where the stock of capital is rising faster than labor productivity, the return to investment should decline with time. For the advanced countries, the evidence does not support this argument because returns have not fallen over time.

Critiques of the neoclassical model led to two lines of subsequent research on economic growth. The first approach, which was originated by Jorgenson (1966, 2000), is termed the augmented Solow approach. It remains in the neoclassical tradition in that diminishing marginal returns are critical and there is no explanation for the determinants of technological progress. Instead, this approach attempts to reduce empirically the portion of growth attributed to the unexplained residual labeled technological progress (TFP). The idea is to develop better measures of the inputs used in the production function and broaden the definition of investment by including human capital, research and development, and public infrastructure. In addition, the composition of capital spending is important. Higher levels of capital spending on high-technology goods will boost productivity more than spending on machine tools or structures.

By adding inputs like human capital to the production function, the augmented Solow model enables us to more accurately measure the contribution of technological progress to growth. However, the economy still moves toward a steady state growth path because even broadly defined capital is assumed to eventually encounter diminishing marginal returns. In essence, this line of research uses the growth accounting methodology and increases the number of inputs in the production function in order to provide a more accurate measure of technological progress. The second approach is the endogenous growth theory, which we examine in the next section.

5.3. Endogenous Growth Theory

The alternative to the neoclassical model is a series of models known as endogenous growth theory. These models focus on explaining technological progress rather than treating it as exogenous. In these models, self-sustaining growth emerges as a natural consequence of the model, and the economy does not necessarily converge to a steady state rate of growth. Unlike the neoclassical model, there are *no diminishing marginal returns to capital for the economy as a whole* in the endogenous growth models. So increasing the saving rate permanently increases the rate of economic growth. These models also allow for the possibility of increasing returns to scale.

Romer (1986) provided a model of technological progress and a rationale for why capital does not experience diminishing marginal returns. He argued that capital accumulation is the main factor accounting for long-run growth, once the definition of capital is broadened to include such items as human or knowledge capital and research and development (R&D). R&D is defined as investment in new knowledge that improves the production process. In endogenous growth theory, knowledge or human capital and R&D spending are factors of production, like capital and labor, and have to be paid for through savings.

Companies spend on R&D for the same reason they invest in new equipment and build new factories: to make a profit. R&D spending is successful if it leads to the development of a new product or method of production that is successful in the marketplace. However, there is a fundamental difference between spending on new equipment and factories and spending on R&D. The final product of R&D spending is usually ideas. These ideas can potentially be copied and used by other companies in the economy. Thus, R&D expenditures have potentially large positive externalities or spillover effects. This means that spending by one company has a positive impact on other companies and increases the overall pool of knowledge available to all companies. Spending by companies on R&D and knowledge capital generates benefits to the economy as a whole that exceed the private benefit to the individual company making the R&D investment. Individual companies cannot fully capture all the benefits associated with creating new ideas and methods of production. Some of the benefits are external to the company, and so are the social returns associated with the investment in R&D and human capital.

This distinction between the private and social returns or benefits to capital is important because it solves an important microeconomic issue. The elimination of the assumption of diminishing marginal returns to capital implies constant returns to capital and increasing returns to all factors taken together. If individual companies could capture these scale economies, then all industries would eventually be dominated by a single company—a monopoly. There is simply no empirical evidence to support this implication. Separating private returns from social returns solves the problem. If companies face constant returns to scale for all private factors, there is no longer an inherent advantage for a company to be large. But the externality or social benefit results in increasing returns to scale across the entire economy as companies benefit from the private spending of the other companies.

The role of R&D spending and the positive externalities associated with this spending have important implications for economic growth. In the endogenous growth model, the economy does not reach a steady growth rate equal to the growth of labor plus an exogenous rate of labor productivity growth. Instead, saving and investment decisions can generate self-sustaining growth at a permanently higher rate. This situation is in sharp contrast to the neoclassical model, in which only a transitory increase in growth above the steady state is possible. The reason for this difference is that because of the externalities on R&D, diminishing marginal returns to capital do not set in. The production function in the endogenous growth model is a straight line given by:

$$y_e = f(k_e) = ck_e \tag{11-12}$$

where output per worker (y_e) is proportional to the stock of capital per worker (k_e), c is the (constant) marginal product of capital in the aggregate economy, and the subscript e denotes the endogenous growth model. In contrast, the neoclassical production function is a curved line that eventually flattens out (see Exhibit 11-4).

To understand the significance of introducing constant returns to aggregate capital accumulation, note that in this model the output-to-capital ratio is fixed ($= c$) and therefore output per worker (y_e) always grows at the same rate as capital per worker (k_e). Thus, faster or slower capital accumulation translates one for one into faster or slower growth in output per capita. Substituting Equation 11-12 into Equation 11-7 gives an equation for the growth rate of output per capita in the endogenous growth model:

$$\Delta y_e/y_e = \Delta k_e/k_e = sc - \delta - n$$

Because all the terms on the right-hand side of this equation are constant, this is both the long-run and short-run growth rate in this model. Examination of the equation shows that *a higher saving rate* (s) implies *a permanently higher growth rate*. This is the key result of the endogenous growth model.

The positive externalities associated with spending on R&D and knowledge capital suggest that spending by private companies on these inputs may be too low from an overall societal point of view. This is an example of a market failure where private companies underinvest in the production of these goods. In this case, there may be a role for government intervention to correct for the market failure by direct government spending on R&D or providing tax breaks and subsidies for private production of knowledge capital. Higher levels of spending on knowledge capital could translate into faster economic growth even in the long run. Finally, according to the endogenous growth theory, there is *no reason why the incomes of developed and developing countries should converge*. Because of constant or even increasing returns associated with investment in knowledge capital, the developed countries can continue to grow as fast as, or faster than, the developing countries. As a result, there is no reason to expect convergence of income over time. We now turn to the convergence debate in more detail.

EXAMPLE 11-12 Neoclassical versus Endogenous Growth Models

Consider again an economy with per capita income growing at a constant 3.08 percent rate and with a 20 percent saving rate, an output-to-capital ratio (c in the endogenous growth model, Equation 11-12) of 0.7040, a depreciation rate (δ) of 10 percent, and a 1 percent labor force growth (n).

1. Use the endogenous growth model to calculate the new steady state growth rate of per capita income if the saving rate increases to 23.5 percent.
2. How much higher will per capita income be in 10 years because of the higher saving rate? How does this compare with the impact calculated in Example 11-11 using the neoclassical model? What accounts for the difference?
3. In an effort to boost growth, the government is considering two proposals. One would subsidize all private companies that increase their investment spending. The second would subsidize only investments in R&D and/or implementation of new technologies with potential for network externalities. Interpret these proposals in terms of the neoclassical and endogenous growth models and assess their likely impact on growth. (Focus only on supply-side considerations here.)

Solution to 1: In the endogenous growth model the new growth rate of per capital income is:

$$\Delta y_e/y_e = sc - \delta - n = (0.235)(0.7040) - 0.10 - 0.01 = 0.0554, \text{ or } 5.54\%$$

This is the same as the growth rate immediately following the increase in the saving rate (to 30 percent in that case) in the earlier example using the neoclassical model

(Example 11-11). Unlike in the neoclassical model, in the endogenous growth model this higher growth rate will be sustained.

Solution to 2: According to the endogenous growth model, per capita income will grow 2.46 percent (= 5.54% – 3.08%) faster with the higher saving rate. After 10 years, the cumulative impact of the faster growth rate will be:

$$\exp(0.0246 \times 10) = \exp(0.246) = 1.2789$$

So, per capita income will be almost 28 percent higher than it would have been at the lower saving rate. This increase is substantially larger than the 15.18 percent cumulative increase after 10 years found in Example 11-11 assuming a much larger increase in the saving rate (to 30 percent instead of 23.5 percent) in the neoclassical model. The difference arises because the endogenous growth model assumes that capital accumulation is not subject to diminishing returns. Therefore, the growth rate is permanently, rather than temporarily, higher.

Solution to 3: Subsidizing all private investment would tend to have a significant, pure capital deepening component. That is, companies would be encouraged to buy more, but not necessarily better, plant and equipment. The neoclassical model indicates that this is likely to result in a temporary surge in growth, but even if the higher rate of investment/saving is sustained, growth will again decline over time. On the positive side, this proposal is very likely to succeed, at least for a while, because it does not require investment in unproven technologies or ill-defined network effects. The impact of the other proposal is more uncertain but potentially much more powerful. If the investments in R&D and/or new technologies lead to new knowledge, greater efficiency, new products and methods, and/or network externalities, then the endogenous growth model suggests that growth is likely to be permanently enhanced.

5.4. Convergence Debate

As is evident in Exhibit 11-1, a wide gap separates the living standards in the developed and developing nations of the world. The question is: Will this difference persist forever or will the per capita income levels of the developing countries converge to those of the developed countries? Convergence means that countries with low per capita incomes should grow at a faster rate than countries with high per capita incomes. Thus, over time the per capita incomes of developing countries should converge toward those of the developed countries. Whether convergence occurs has major implications for the future growth prospects of developed versus developing countries. It also has important investment implications.

Neoclassical growth theory predicts two types of convergence: absolute convergence and conditional convergence. **Absolute convergence** means that developing countries, regardless of their particular characteristics, will eventually catch up with the developed countries and match them in per capita output. The neoclassical model assumes that all countries have access to the same technology. As a result, per capita income in all countries should eventually grow at the same rate. Thus, the model implies convergence of per capita *growth rates* among all

countries. It does not, however, imply that the *level* of per capita income will be the same in all countries regardless of underlying characteristics; that is, it does not imply absolute convergence.

Conditional convergence means that convergence is conditional on the countries having the same saving rate, population growth rate, and production function. If these conditions hold, the neoclassical model implies convergence to the same *level* of per capita output as well as the same steady state growth rate. In terms of Exhibit 11-13, these economies would have the same k^* and thus the same steady state. If they start with different capital-to-labor ratios, their growth rates will differ in the transition to the steady state. The economy with a lower capital-to-labor ratio will experience more rapid growth of productivity and per capita income, but the differential will diminish until they finally converge. Countries with different saving rates or population growth rates and thus different steady state values for k^* will have different steady state *levels* of per capita income, but their growth rates of per capita output will still converge.

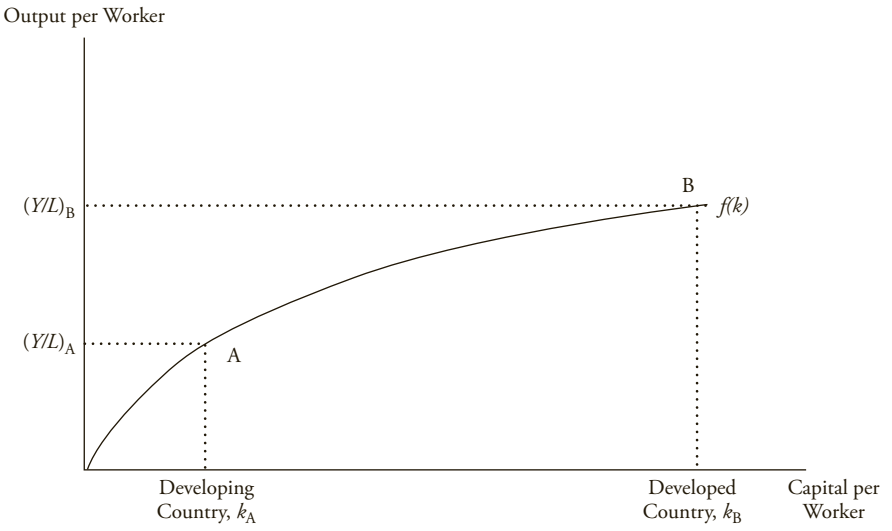
The data (see Exhibit 11-18 later in the chapter) indicate that some of the poorer countries are diverging rather than converging to the income levels of the developed countries. Thus, in addition to the first two convergence concepts, we have the notion of **club convergence**, where only rich and middle-income countries that are members of the club are converging to the income level of the richest countries in the world. This means that the countries with the lowest per capita income in the club grow at the fastest rate. In contrast, countries outside the club continue to fall behind. Poor countries can join the club if they make appropriate institutional changes, such as those summarized in Section 2.7. Finally, countries may fall into a **nonconvergence trap** if they do not implement necessary institutional reforms. For example, failure to reform labor markets has undermined growth in some European countries that have experienced weak growth in employment and high rates of unemployment over the past two decades. Certain institutional arrangements that initially enhance growth may later generate nonconvergence traps if maintained too long. Import substitution policies enabled the Latin American countries to grow rapidly in the 1950s and 1960s but caused them to stagnate in the 1970s and 1980s.

If convergence, and especially club convergence, does occur, investing in countries with lower per capita incomes that are members of the club should, over long periods of time, provide a higher rate of return than investing in higher-income countries. Convergence means that the rate of growth of potential GDP should be higher in developing countries that have made the institutional changes that are a precondition for growth and that enable these countries to become members of the convergence club. With higher long-term growth in these economies, corporate profits should also grow at a faster rate. Given the faster rate of growth in earnings, stock prices may also rise at a faster rate. Of course, risk is also likely to be higher in these markets. Nonetheless, it is reasonable to conclude that long-term investors should allocate a risk-tolerance-appropriate portion of their assets to those developing economies that have become members of the convergence club.

Convergence between the developed and developing countries can occur in two ways. First, convergence takes place through capital accumulation and capital deepening. Exhibit 11-17 illustrates the difference between developed and developing countries using the per capita neoclassical production function. The developed countries operate at point B, so increases in capital have almost no impact on productivity. In contrast, developing countries operate at point A, where increases in capital significantly boost labor productivity.

A second source of convergence is that developing countries can imitate or adopt technology already widely utilized in the advanced countries. Developing countries can learn from

EXHIBIT 11-17 Per Capita Production Function, Developed versus Developing Countries



advanced countries as scientific and management practices spread with globalization. By importing technology from the advanced countries, the developing countries can achieve faster economic growth and converge to the income of the advanced countries. Japan had a successful imitation strategy in the 1950s and 1960s, as does China now. Technology transfers will narrow the income gap between developed and developing countries only if the poor countries invest the resources to master the technology and apply it to their economies. This spending is similar to R&D spending and allows the country to join the convergence club. The steady state rate of growth for members of the convergence club will be determined by the global rate of technological progress. Without such spending, the country will be left out and will continue to fall behind the developed countries.

In contrast to the neoclassical model, the endogenous growth model makes no prediction that convergence should occur. This model allows for countries that start with high per capita income and more capital to grow faster and stay ahead of the developing countries. If the externalities associated with knowledge and human capital are large, the higher-income country can maintain its lead through high rates of investment in these capital inputs.

If the convergence hypothesis is correct, there should be an inverse relationship between the initial level of per capita real GDP and the growth rate in per capita GDP. Exhibit 11-18 shows the countries in Exhibit 11-1 (except Germany) in descending order of per capita income in 1950. If incomes are converging across countries, the poor countries in 1950 should have a higher growth rate between 1950 and 2010 than the rich countries.

The results for the convergence hypothesis are mixed. The countries with the highest per capita incomes in 1950 were the United States, New Zealand, Australia, and Canada. The markets with the fastest growth rates over the period 1950–2010 were Taiwan, South Korea, and China, each growing at a rate above 5 percent. This result strongly supports convergence because the per capita incomes of all three countries in 1950 were well below that of the United States. In addition, the results for Japan, Singapore, and Spain showed a convergence

EXHIBIT 11-18 Real Per Capita GDP by Country

	Real GDP Per Capita in Dollars				Average Annual Growth in Real Per Capita GDP (%)			
	1950	1970	1990	2010	1950–1970	1970–1990	1990–2010	1950–2010
United States	14,559	22,806	35,328	46,697	2.27	2.21	1.40	1.96
New Zealand	13,795	18,255	22,331	31,223	1.41	1.01	1.69	1.37
Australia	13,219	21,444	30,628	45,951	2.45	1.80	2.05	2.10
Canada	12,053	19,919	31,196	41,288	2.54	2.27	1.41	2.07
United Kingdom	11,602	18,002	27,469	37,378	2.22	2.14	1.55	1.97
France	8,266	18,186	28,127	34,358	4.02	2.20	1.01	2.40
Venezuela	8,104	11,590	9,028	10,560	1.81	−1.24	0.79	0.44
Argentina	6,164	9,026	7,952	13,468	1.93	−0.63	2.67	1.31
Italy	5,954	16,522	27,734	31,069	5.24	2.62	0.57	2.79
Ireland	5,496	9,869	18,812	36,433	2.97	3.28	3.36	3.20
Saudi Arabia	5,060	17,292	20,399	22,951	6.34	0.83	0.59	2.55
South Africa	4,361	6,959	6,595	8,716	2.36	−0.27	1.40	1.16
Singapore	4,299	8,600	27,550	56,224	3.53	5.99	3.63	4.38
Mexico	4,180	7,634	10,754	13,710	3.06	1.73	1.22	2.00
Spain	3,964	11,444	21,830	30,504	5.44	3.28	1.69	3.46
Peru	3,464	5,786	4,516	8,671	2.60	−1.23	3.32	1.54
Hong Kong	3,128	8,031	24,734	43,324	4.83	5.79	2.84	4.48
Japan	3,048	15,413	29,813	34,828	8.44	3.35	0.78	4.14
Brazil	2,365	4,324	6,959	9,589	3.06	2.41	1.62	2.36
Turkey	2,327	4,413	7,741	11,769	3.25	2.85	2.12	2.74
Taiwan	1,425	3,948	15,465	36,413	5.23	7.07	4.37	5.55
Philippines	1,296	2,136	2,660	3,672	2.53	1.10	1.63	1.75
South Korea	1,185	3,009	12,083	30,079	4.77	7.20	4.67	5.54
Egypt	1,132	1,560	3,137	5,306	1.62	3.55	2.66	2.61
Nigeria	814	1,183	1,203	2,037	1.89	0.08	2.67	1.54
Indonesia	804	1,182	2,517	4,740	1.95	3.85	3.22	3.00
Kenya	791	1,113	1,359	1,376	1.72	1.00	0.06	0.93
Vietnam	689	770	1,073	3,369	0.56	1.67	5.89	2.68

(*Continued*)

EXHIBIT 11-18 *Continued*

	Real GDP Per Capita in Dollars				Average Annual Growth in Real Per Capita GDP (%)			
	1950	1970	1990	2010	1950–1970	1970–1990	1990–2010	1950–2010
Pakistan	666	985	1,645	2,600	1.98	2.60	2.32	2.29
India	658	922	1,390	3,575	1.70	2.07	4.84	2.86
Botswana	449	774	3,731	5,311	2.76	8.18	1.78	4.20
China	402	698	1,677	8,569	2.80	4.48	8.50	5.23
Ethiopia	314	479	462	749	2.13	−0.18	2.45	1.46

to the level of income in the advanced economies. In total, 23 of the 32 countries grew faster than the United States over the period. However, Ethiopia, Kenya, Nigeria, the Philippines, Peru, South Africa, Argentina, Venezuela, and New Zealand fell further behind the United States. Interestingly, since 1990 convergence has been relatively strong overall, with 24 countries (75 percent) growing faster than the United States—including Ethiopia, Nigeria, the Philippines, Peru, South Africa, Argentina, and New Zealand—but has not continued among the most advanced economies, as France, Japan, and Italy lagged the United States, Canada, and Australia.

The evidence seems to suggest that poorer countries may converge if they develop the appropriate legal, political, and economic institutions as discussed in Section 2.7. In addition, trade policy is an important factor, which we address in the next section.

6. GROWTH IN AN OPEN ECONOMY

The Solow model discussed in Section 5.2 assumed a closed economy in which domestic investment equals domestic savings and there is no international trade or capital flows. Opening up the economy to trade and financial flows can significantly affect the rate of growth in an economy for the following five reasons:

1. A country can borrow or lend funds in global markets, and domestic investment can be funded by global savings. Thus, investment is not constrained by domestic savings.
2. Countries can shift resources into industries in which they have a comparative advantage and away from industries in which they are relatively inefficient, thereby increasing overall productivity.
3. Companies have access to a larger, global market for their products, allowing them to better exploit any economies of scale and increasing the potential reward for successful innovation.
4. Countries can import technology, thus increasing the rate of technological progress.
5. Global trade increases competition in the domestic market, forcing companies to produce better products, improve productivity, and keep costs low.

According to the neoclassical model, convergence should occur more quickly if economies are open and there is free trade and international borrowing and lending. Opening up the economy should increase the rate at which countries' capital-to-labor ratios converge. The dynamic adjustment process can be described in seven parts:

1. Developing countries have less capital per worker, and as a result, the marginal product of capital is higher. Thus, the rate of return on investments should be higher in countries with low capital-to-labor ratios and lower in countries with high capital-to-labor ratios.
2. Global savers, seeking higher returns on investments, will invest in the capital-poor countries. In an open economy, capital should flow from countries with high capital-to-labor ratios to those that are capital poor.
3. Because of the capital inflows, the physical capital stock in the developing countries should grow more rapidly than in rich countries even if the saving rate is low in the poorer countries. Faster capital growth will result in higher productivity growth, causing per capita incomes to converge.
4. Because capital flows must be matched by offsetting trade flows, capital-poor countries will tend to run a trade deficit as they borrow globally to finance domestic investment. In contrast, the developed countries will tend to run trade surpluses as they export capital.
5. During the transition to the new steady state, the inflows of capital will temporarily raise the rate of growth in the capital-poor country above the steady state rate of growth. At the same time, growth in the capital-exporting countries will be below the steady state.
6. Over time, the physical capital stock will rise in the capital-poor country, reducing the return on investments. As a result, the rate of investment and size of the country's trade deficit will decline. Growth will slow and approach the steady state rate of growth. If investment falls below the level of domestic savings, the country will eventually shift from a trade deficit to a trade surplus and become an exporter of capital.
7. In the Solow model, after the reallocation of world savings, there is no permanent increase in the rate of growth in an economy. Both the developed and developing countries grow at the steady state rate of growth.

In contrast to the neoclassical model, endogenous growth models predict that a more open trade policy will permanently raise the rate of economic growth. In these models, international trade increases global output through the following three effects:

1. A selection effect, where increased competition from foreign companies forces less efficient domestic companies to exit and more efficient ones to innovate, raising the efficiency of the overall national economy.
2. A scale effect that allows producers to more fully exploit economies of scale by selling to a larger market.
3. A backwardness effect arising from less advanced countries or sectors of an economy catching up with the more advanced countries or sectors through knowledge spillovers.

Open trade also affects the innovation process by encouraging higher levels of spending on R&D and on human capital as companies invest to take advantage of access to larger markets and the greater flow of ideas and knowledge among countries. The rate of return to new investment increases, as does the rate of economic growth. In general, most countries gain from open trade, with the scale effect benefiting smaller countries and the backwardness effect benefiting the poorer, less developed countries. But trade can also retard growth in some cases,

especially in small countries that lag behind the technology leaders. Opening these countries to trade may discourage domestic innovation because companies will recognize that, even if they innovate, they may lose out to more efficient foreign companies.

The Entry of China and India into the Global Economy

China and India effectively entered the global economy in the 1980s as they shifted toward more market-oriented policies and opened up to global trade. Their impact on global growth was significant. In 2010, China and India accounted for 13.6 percent and 5.5 percent of world GDP, respectively, whereas the two countries had combined for only 4.2 percent of global output in 1980. The entry of these two countries significantly increased the global supply of skilled and unskilled labor receiving low wages. As a result of the surge in available labor, global potential GDP increased sharply. Economic theory suggests that the supply-side increase in the global capacity to produce goods and services would increase global output and put downward pressure on prices.

The neoclassical model of growth can provide us with some further insights into the impact of China and India entering the global economy. China and India are low-wage and capital-poor countries. In contrast, the United States, Japan, and Europe are high-wage and capital-rich countries. One would expect that the rate of return on capital would be higher in China and India and that capital would flow from the developed countries to China and India. Hence, both China and India would be expected to run trade deficits. This has been the case for India but, contrary to the prediction of the model, China has run trade surpluses. These surpluses stem from China's very high domestic saving rate—even higher than its high investment rate—as well as currency intervention that channels domestic saving into accumulation of foreign exchange reserves.

Nonetheless, China has experienced large foreign direct investment inflows (see Exhibit 11-19), which have reinforced its already high private investment rate. As China and India accumulate capital, their capital-to-labor ratios, real wage levels, and per capita income should converge toward those of the advanced economies. Depending on global aggregate demand conditions, wages might even have to fall in the developed countries in the process of shifting wealth and income to the developing economies. Because of the surge in the global supply of labor, the overall share of labor in global income should decline relative to capital. In addition, global productivity should rise as China and India account for a rising share of global output. In sum, over the long run, the growing share of global GDP going to China and India will benefit the global economy as more efficient utilization of resources allows global potential GDP to grow more rapidly for an extended period.

Although both the neoclassical and endogenous models of growth show the benefits of open markets, over the past 50 years developing countries have pursued two contrasting strategies for economic development:

1. *Inward-oriented policies* attempt to develop domestic industries by restricting imports. Instead of importing goods and services, these policies encourage the production of

domestic substitutes, despite the fact that it may be more costly to do so. These policies are also called import substitution policies.
2. *Outward-oriented policies* attempt to integrate domestic industries with those of the global economy through trade and to make exports a key driver of growth.

Many African and Latin American countries pursued inward-oriented policies from the 1950s to the 1980s that resulted in poor GDP growth and inefficient industries producing low-quality goods. In contrast, many East Asian countries, such as Hong Kong, Singapore, and South Korea, pursued outward-oriented polices during this same time period, which resulted in high rates of GDP growth and convergence with developed countries. These countries also benefited from the positive effects of foreign direct investment, which suggests that more open and trade-oriented economies will grow at a faster rate. The evidence strongly supports this case.

In Example 11-1, we compared the economic performances of Argentina and Venezuela with those of Japan, South Korea, and Singapore. In 1950, the per capita GDPs of the two Latin American countries were well above those of the three East Asian countries. By 2010, however, the per capita GDPs of the three Asian countries were well above those of Argentina and Venezuela. The difference in the growth rates between Argentina and Venezuela and the three Asian countries is explained largely by the degree of openness of their economies. Argentina and Venezuela were relatively closed economies, whereas the Asian countries relied on foreign investment and open markets to fuel growth. China is now using the same approach and has achieved significant success. China's real per capita GDP (see Exhibit 11-1) increased from $698 in 1970 to $8,569 in 2010 following policy changes that opened the economy to global trade.

The good news is that many African and Latin American countries have removed trade barriers and are now pursuing more outward-oriented policies. These countries have seen better growth in recent years. Brazil is a good example. Exports of goods and services increased from $64.6 billion in 2000 to $249.9 billion in 2010, an increase of over 286 percent. As shown in Exhibit 11-19, exports as a share of GDP rose from 5.1 percent to 10.7 percent over this period.

EXHIBIT 11-19 Exports and Foreign Direct Investment of Selected Countries

	1980	1990	2000	2010
Brazil				
Exports as percentage of GDP	5.1	4.5	5.2	10.7
Inflows of foreign direct investment ($ billions)	NA	NA	$32.8	$48.5
China				
Exports as percentage of GDP	5.9	7.5	9.3	17.2
Inflows of foreign direct investment ($ billions)	NA	NA	$38.4	$185.0
India				
Exports as percentage of GDP	2.7	3.5	3.8	7.2
Inflows of foreign direct investment ($ billions)	NA	NA	$3.6	$24.6

(*Continued*)

EXHIBIT 11-19 *Continued*

	1980	1990	2000	2010
Ireland				
Exports as percentage of GDP	32.1	59.6	88.3	128.5
Inflows of foreign direct investment ($ billions)	NA	NA	$25.8	$26.3
Mexico				
Exports as percentage of GDP	12.8	16.2	15.7	19.1
Inflows of foreign direct investment ($ billions)	NA	NA	$18.0	$19.6
South Africa				
Exports as percentage of GDP	NA	11.6	10.7	16.1
Inflows of foreign direct investment ($ billions)	NA	NA	$0.9	$1.56
South Korea				
Exports as percentage of GDP	NA	26.3	25.1	35.0
Inflows of foreign direct investment ($ billions)	NA	NA	$9.3	−$0.2
United States				
Exports as percentage of GDP	8.8	9.1	10.9	12.7
Inflows of foreign direct investment ($ billions)	NA	NA	$159.2	$351.3

Source: OECD StatLink.

EXAMPLE 11-13 Why Some Countries Converge and Others Do Not

As evident from the high rates of growth between 1950 and 2010 shown in Exhibit 11-18, China and South Korea are converging toward the income levels of the advanced countries but still have a long way to go, especially in the case of China. In contrast, the economies of Mexico and South Africa have not converged toward those of advance countries. Using the data in Exhibits 11-9 and 11-19, give some reasons why this has occurred.

Solution: Two reasons largely account for the difference. First, growth in the Chinese and South Korean economies has been driven by high rates of business investment. As shown in Exhibit 11-9, investment as a share of GDP in 2010 was 48.2 percent in China, almost double the rate of 25.0 percent in Mexico and more than double the rate of 19.3 percent in South Africa. Although investment as a share of GDP in South Korea is lower than in China, it is well above that of Mexico and South Africa.

Second, both China and South Korea have pursued an aggressive export-driven, outward-oriented policy focusing on manufactured goods. In 2010, exports were 35 percent of GDP for South Korea and 17.2 percent of GDP for China (Exhibit 11-19). In addition, foreign direct investment is a major factor underlying growth in China.

The comparable export numbers for Mexico and South Africa are 19.1 percent and 16.1 percent of GDP, respectively. Despite the North American Free Trade Agreement (NAFTA), Mexico's exports as a share of GDP rose only modestly from 1990 to 2010. In contrast, exports as a share of GDP for China have nearly doubled since 2000. In addition, Mexico and South Africa attracted only a combined $21.2 billion in foreign direct investment in 2010, significantly less than that of Ireland—a smaller but much wealthier and very open country—and the $185 billion inflow of foreign investment into China. The upshot is that Mexico and South Africa have been more inward-oriented economies. These trends are changing, however, as many African and Latin American countries are increasingly relying on growing exports and foreign investment to increase GDP growth.

EXAMPLE 11-14 Investment Prospects in Spain: Estimating the Sustainable Growth Rate

You are a financial analyst at Global Invest Inc., an investment management firm that runs a number of global mutual funds with a significant exposure to Spain. The Madrid General Index, which reached a crisis-induced low of 716 in March 2009, remains far below its November 2007 peak of 1,725. The members of the investment policy committee at the firm believe the equity market in Spain is attractive and is currently being depressed by temporary problems in the banking and real estate markets of Spain, which they feel are overstated. They believe that higher profits will ultimately drive the market higher but are concerned about the long-term prospects and the sustainable rate of growth for Spain. One of the research assistants at the firm gathers the data shown in Exhibit 11-20 from the OECD and the Conference Board.

EXHIBIT 11-20 Growth Data for Spain

	GDP in Billions of USD Adjusted for PPP	Gross Capital Spending as Percentage of GDP	Consumption of Fixed Capital as Percentage of GDP	Labor Cost as Percentage of Total Factor Cost	Total Hours Worked (millions)	Output per Hour Worked in 2009 USD Adjusted for PPP	Growth in Total Factor Productivity (%)
2000	1,156.4	26.3	13.6	64.40	28,402	40.7	−0.87
2001	1,198.6	26.4	13.7	63.66	29,232	41.0	−0.78
2002	1,231.0	26.6	14.1	62.97	29,836	41.3	−0.57
2003	1,269.1	27.4	14.4	62.69	30,495	41.6	−0.21
2004	1,310.6	28.3	14.9	61.71	31,274	41.9	−0.58
2005	1,357.9	29.5	15.3	60.87	32,132	42.3	−0.65
2006	1,412.5	31.0	15.6	60.66	33,146	42.6	−0.28

(*Continued*)

EXHIBIT 11-20 *Continued*

	GDP in Billions of USD Adjusted for PPP	Gross Capital Spending as Percentage of GDP	Consumption of Fixed Capital as Percentage of GDP	Labor Cost as Percentage of Total Factor Cost	Total Hours Worked (millions)	Output per Hour Worked in 2009 USD Adjusted for PPP	Growth in Total Factor Productivity (%)
2007	1,462.9	30.9	15.7	60.04	33,757	43.3	−0.07
2008	1,475.6	29.1	16.2	60.23	33,830	43.6	−1.63
2009	1,420.6	24.4	16.9	59.47	31,705	44.8	−1.61

Sources: OECD.StatExtracts and the Conference Board Total Economy Database.

From the Conference Board website, the physical capital stock for Spain was estimated at $2,177.2 billion (adjusted for purchasing power parity) in 1999. The research analyst calculated the physical capital stock (K) for Spain for the years 2000–2009 using the following equation:

$$K_t = K_{t-1} + I - D$$

where I is gross investment or gross capital spending and D is the depreciation or the consumption of fixed capital. So for 2000 and 2001, the physical capital stock is calculated as:

$$K_{2000} = \$2{,}177.2 + \$1{,}156.4(0.263 - 0.136) = \$2{,}324.1 \text{ billion}$$
$$K_{2001} = \$2{,}324.1 + \$1{,}198.6(0.264 - 0.317) = \$2{,}476.3 \text{ billion}$$

The physical capital stock for the remaining years is calculated in the same way and given by Exhibit 11-21.

EXHIBIT 11-21 Estimated Physical Capital Stock (USD billions)

2000	2,324.1
2001	2,476.3
2002	2,630.2
2003	2,795.1
2004	2,970.8
2005	3,163.6
2006	3,381.1
2007	3,603.5
2008	3,793.8
2009	3,900.4

You are requested by the investment policy committee to use this data to address the following:

1. Calculate the potential growth rate of the Spanish economy using the production function or growth accounting method (Equation 11-4), and determine the amount of growth attributed to each source.
2. Calculate the potential growth rate of the Spanish economy using the labor productivity method (Equation 11-5).
3. How significant are capital deepening and technology in explaining growth for Spain?
4. What is the steady state growth rate for Spain according to the neoclassical model?
5. Assess the implications of the growth analysis for future economic growth and equity prices in Spain.

Solution to 1: The production function or growth accounting method estimates the growth in GDP using Equation 11-4:

$$\text{Growth in potential GDP} = \alpha\Delta K/K + (1-\alpha)\Delta L/L + \Delta A/A$$

The annual growth rate in capital is calculated from Exhibit 11-21 as:*

$$(3{,}900.4/2{,}324.1)^{1/9} - 1 = 5.92\%$$

The labor input is measured by the growth rate in total hours worked in the economy (Exhibit 11-20) and given by:

$$(31{,}705/28{,}402)^{1/9} - 1 = 1.23\%$$

The growth rate in total factor productivity (Exhibit 11-20) is calculated by using a geometric average of the growth rates for 2000–2009 and is equal to −0.73 percent. Finally, the labor share of output is given by the average of the labor cost as a percentage of total factor cost, which is 61.7 percent for 2000–2009 (Exhibit 11-20). Thus, the share of capital (α) is 1 – 0.617 = 38.3%.

Using these numbers, the growth in potential GDP is:

$$\begin{aligned}\text{Growth in potential GDP} &= \alpha\Delta K/K + (1-\alpha)\Delta L/L + \Delta A/A \\ &= (0.383)0.0592 + (0.617)0.0123 + (-0.0073) = 2.30\%\end{aligned}$$

Sources of growth for Spain over the period 2000–2009 were:

Capital	$0.383 \times 0.0592 = 2.27\%$
Labor	$0.617 \times 0.0123 = 0.76\%$
TFP	$= -0.73\%$

Solution to 2: The labor productivity method estimates the growth in GDP using Equation 11-5:

$$\text{Growth rate in potential GDP} = \text{Long-term growth rate of labor force} + \text{Long-term growth rate of in labor productivity}$$

As before, we use the growth in total hours worked to measure the growth in the labor force. The growth in labor productivity per hour worked is:

$$(44.8/40.7)^{1/9} - 1 = 1.07\%$$
$$\text{Growth in potential GDP} = 1.23\% + 1.07\% = 2.3\%$$

Note that the estimate of potential GDP growth using the labor productivity approach is the same as that obtained from the growth accounting method. In general, the two methods are likely to give somewhat different estimates because they rely on different data inputs. The growth accounting method requires measurements of the physical capital stock and TFP. As discussed in Section 4.3, TFP is estimated using various time-series or econometric models of the component of growth that is not accounted for by the explicit factors of production. As a result, the estimate of TFP reflects the average (or smoothed) behavior of the growth accounting residual. The labor productivity approach is simpler, and it avoids the need to estimate the capital input and TFP. In contrast to the estimated value of TFP, labor productivity is measured as a pure residual; that is, it is the part of GDP growth that is not explained by the labor input (and only the labor input). The cost of the simplification is that the labor productivity approach does not allow a detailed analysis of the drivers of productivity growth.

Solution to 3: Capital deepening occurs in an economy when there is an increase in the capital-to-labor ratio. The labor input for Spain is measured in terms of total hours worked in the economy. Thus, the capital-to-labor ratio for Spain is calculated by dividing the physical capital stock in Exhibit 11-21 by total hours worked in Exhibit 11-20. The results, shown in Exhibit 11-22, indicate that capital deepening was very significant in Spain: The amount of capital per hour worked increased from \$81.83 in 2000 to \$123.02 in 2009. In terms of the growth rate, the capital-to-labor ratio increased at an annual rate of 4.6 percent.

The contribution of technology is measured by the growth in total factor productivity (TFP). In contrast to capital deepening, TFP made a negative contribution to growth; the average rate of growth for TFP from 2000 to 2009 was −0.73 percent. However, TFP is estimated using various statistical techniques, and given the uncertainty around these estimates, it should be viewed with some caution.

Solution to 4: The steady state growth rate in the neoclassical model is estimated by (see Equation 11-8):

$$\Delta Y/Y = (\theta)/(1-\alpha) + n = \text{Growth rate of TFP scaled by labor factor share} + \text{Growth rate in the labor force}$$
$$\text{Steady state growth rate} = -0.73\%/(1-0.383) + 1.23\% = 0.05\%$$

EXHIBIT 11-22 Estimated Capital-to-Labor Ratio ($ millions/hour worked)

2000	$81.83
2001	84.71
2002	88.16
2003	91.66
2004	94.99
2005	98.46
2006	102.01
2007	106.75
2008	112.14
2009	123.02

As expected, the growth rate in potential GDP (calculated as in the solutions to Questions 1 and 2) is above the steady state growth rate. The reason for this is that the economy of Spain is still in the process of converging to the higher income levels of the United States and the major economies in Europe. The physical capital stock is below the steady state, and capital deepening is a significant factor increasing productivity growth and the growth in potential GDP. Steady state growth may be somewhat underestimated in our analysis given that TFP growth is likely to revert to the 1 percent annual rate of increase exhibited in other major developed economies. This is likely to be offset by a lower growth rate in the labor input (see Example 11-6).

Solution to 5: The results suggest that potential GDP growth in Spain is approximately 2.3 percent. As we saw in Exhibit 11-1, the growth rate of actual GDP since early 2000 has been 2.1 percent per year, close to the previous estimate of potential but well above the steady state. The problem is that all the growth in potential GDP is due to the increase in the labor and capital inputs, with capital deepening being very significant, as the capital-to-labor ratio is increasing at a 4.6 percent annual rate. The neoclassical model would suggest that the impact of capital deepening will decline over time and the economy will move toward a steady state rate of growth. Thus, growth based on capital deepening should not be sustainable over time. The other major question raised is whether the labor input can continue to grow at an annual rate of 1.2 percent. We examined this question in Example 11-6. In sum, potential GDP growth is likely to fall over time given Spain's reliance on capital deepening and the strong possibility that growth in the labor input is likely to slow. However, the reversion of TFP growth to levels more typical of other European economies should mitigate the decline. Even if TFP does rebound, slower growth in potential GDP in Spain will likely restrain future stock price increases.

*Using the 1999 capital stock as a base instead of the 2000 capital stock would give almost the same growth rate: $(3{,}900.4/2{,}177.2)^{1/10} - 1 = 6.00\%$.

7. SUMMARY

This chapter focuses on the factors that determine the long-term growth trend in the economy. As part of the development of global portfolio equity and fixed-income strategies, investors must be able to determine both the near-term and the sustainable rates of growth within a country. Doing so requires identifying and forecasting the factors that determine the level of GDP and that determine long-term sustainable trends in economic growth.

- The sustainable rate of economic growth is measured by the rate of increase in the economy's productive capacity or potential GDP.
- Growth in real GDP measures how rapidly the total economy is expanding. Per capita GDP, defined as real GDP divided by population, measures the standard of living in each country.
- The growth rate of real GDP and the level of per capita real GDP vary widely among countries. As a result, investment opportunities differ by country.
- Equity markets respond to anticipated growth in earnings. Higher sustainable economic growth should lead to higher earnings growth and equity market valuation ratios, all other things being equal.
- The best estimate for the long-term growth in earnings for a given country is the estimate of the growth rate in potential GDP.
- In the long run, the growth rate of earnings cannot exceed the growth in potential GDP. Labor productivity is critical because it affects the level of the upper limit. A permanent increase in productivity growth will raise the upper limit on earnings growth and should translate into faster long-run earnings growth and a corresponding increase in stock price appreciation.
- For global fixed-income investors, a critical macroeconomic variable is the rate of inflation. One of the best indicators of short- to intermediate-term inflation trends is the difference between the growth rate of actual and potential GDP.
- Capital deepening, an increase in the capital-to-labor ratio, occurs when the growth rate of capital (net investment) exceeds the growth rate of labor. In a graph of output per capita versus the capital-to-labor ratio, it is reflected by a move along the curve (i.e., the production function).
- An increase in total factor productivity (TFP) causes a proportional upward shift in the entire production function.
- One method of measuring sustainable growth uses the production function and the growth accounting framework developed by Solow. It arrives at the growth rate of potential GDP by estimating the growth rates of the economy's capital and labor inputs plus an estimate of total factor productivity.
- An alternative method measures potential growth as the long-term growth rate of the labor force plus the long-term growth rate of labor productivity.
- The forces driving economic growth include the quantity and quality of labor and the supply of non-ICT and ICT capital, public capital, raw materials, and technological knowledge.
- The labor supply is determined by population growth, the labor force participation rate, and net immigration. The physical capital stock in a country increases with net investment. The correlation between long-run economic growth and the rate of investment is high.
- Technological advances are discoveries that make it possible to produce more or higher-quality goods and services with the same resources or inputs. Technology is a major factor

determining TFP. TFP is the main factor affecting long-term, sustainable economic growth rates in developed countries and also includes the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices.

- Total factor productivity, estimated using a growth accounting equation, is the residual component of growth once the weighted contributions of all explicit factors (e.g., labor and capital) are accounted for.
- Labor productivity is defined as output per worker or per hour worked. Growth in labor productivity depends on capital deepening and technological progress.
- The academic growth literature is divided into three theories—the classical view, the neoclassical model, and the new endogenous growth view.
- In the classical model, growth in per capita income is only temporary because an exploding population with limited resources brings per capita income growth to an end.
- In the neoclassical model, a sustained increase in investment increases the economy's growth rate only in the short run. Capital is subject to diminishing marginal returns, so long-run growth depends solely on population growth, progress in TFP, and labor's share of income.
- The neoclassical model assumes that the production function exhibits diminishing marginal productivity with respect to any individual input.
- The point at which capital per worker and output per worker are growing at equal, sustainable rates is called the steady state or balanced growth path for the economy. In the steady state, total output grows at the rate of labor force growth plus the rate of growth of TFP divided by the elasticity of output with respect to labor input.
- The following parameters affect the steady state values for the capital-to-labor ratio and output per worker: saving rate, labor force growth, growth in TFP, depreciation rate, and elasticity of output with respect to capital.
- The main criticism of the neoclassical model is that it provides no quantifiable prediction of the rate or form of TFP change. TFP progress is regarded as exogenous to the model.
- Endogenous growth theory explains technological progress within the model rather than treating it as exogenous. As a result, self-sustaining growth emerges as a natural consequence of the model and the economy does not converge to a steady state rate of growth that is independent of saving/investment decisions.
- Unlike the neoclassical model, where increasing capital will result in diminishing marginal returns, the endogenous growth model allows for the possibility of constant or even increasing returns to capital in the aggregate economy.
- In the endogenous growth model, expenditures made on R&D and for human capital may have large positive externalities or spillover effects. Private spending by companies on knowledge capital generates benefits to the economy as a whole that exceed the private benefit to the company.
- The convergence hypothesis predicts that the rates of growth of productivity and GDP should be higher in the developing countries. Those higher growth rates imply that the per capita GDP gap between developing and developed economies should narrow over time. The evidence on convergence is mixed.
- Countries fail to converge because of low rates of investment and savings, lack of property rights, political instability, poor education and health, restrictions on trade, and tax and regulatory policies that discourage work and investing.
- Opening an economy to financial and trade flows has a major impact on economic growth. The evidence suggests that more open and trade-oriented economies will grow at a faster rate.

PRACTICE PROBLEMS

The following information refers to Questions 1 through 6.[25]

Hans Schmidt, CFA, is a portfolio manager with a boutique investment firm that specializes in sovereign credit analysis. Schmidt's supervisor asks him to develop estimates for GDP growth for three countries. Information on the three countries is provided in Exhibit A.

EXHIBIT A Select Economic Data for Countries A, B, C

Country	Economy	Capital per Worker
A	Developed	High
B	Developed	High
C	Developing	Low

After gathering additional data on the three countries, Schmidt shares his findings with colleague, Sean O'Leary. After reviewing the data, O'Leary notes the following observations:

Observation 1: The stock market of Country A has appreciated considerably over the past several years. Also, the ratio of corporate profits to GDP for Country A has been trending upward over the past several years, and is now well above its historical average.

Observation 2: The government of Country C is working hard to bridge the gap between its standard of living and the living standards of developed countries. Currently, the rate of potential GDP growth in Country C is high.

Schmidt knows that a large part of the analysis of sovereign credit is to develop a thorough understanding of what the potential GDP growth rate is for a particular country and the region in which the country is located. Schmidt is also doing research on Country D for a client of the firm. Selected economic facts on Country D are provided in Exhibit B.

EXHIBIT B Select Economic Facts for Country D

- Slow GDP growth
- Abundant natural resources
- Developed economic institutions

Prior to wrapping up his research, Schmidt schedules a final meeting with O'Leary to see if he can provide any other pertinent information. O'Leary makes the following statements to Schmidt:

[25]These practice problems were developed by Karen Ashby, CFA (LaGrange, Kentucky, USA).

Statement 1: "Many countries that have the same population growth rate, savings rate, and production function will have growth rates that converge over time."

Statement 2: "Convergence between countries can occur more quickly if economies are open and there is free trade and international borrowing and lending; however, there is no permanent increase in the rate of growth in an economy from a more open trade policy."

1. Based on Exhibit A, the factor that would *most likely* have the greatest positive impact on the per capita GDP growth of Country A is:
   A. free trade.
   B. technology.
   C. saving and investment.

2. Based on Observation 1, in the long run the ratio of profits to GDP in Country A is *most likely* to:
   A. remain near its current level.
   B. increase from its current level.
   C. decrease from its current level.

3. Based on Observation 2, Country C is *most likely* to have:
   A. relatively low real asset returns.
   B. a relatively low real interest rate.
   C. a relatively high real interest rate.

4. Based on Exhibit B, the *least likely* reason for the current pace of GDP growth in Country D is:
   A. a persistently strong currency.
   B. strong manufacturing exports.
   C. strong natural resource exports.

5. The type of convergence described by O'Leary in Statement 1 is *best* described as:
   A. club convergence.
   B. absolute convergence.
   C. conditional convergence.

6. Which of the following growth models is *most* consistent with O'Leary's Statement 2?
   A. Classical
   B. Endogenous
   C. Neoclassical

The following information relates to Questions 7 through 15.[26]

Victor Klymchuk, the chief economist at ECONO Consulting (EC), is reviewing the long-term GDP growth of three countries over the recent decade. Klymchuk is interested in forecasting the long-term change in stock market value for each country. Exhibit C presents

[26]These practice problems were developed by Lou Lemos, CFA (Louisville, Kentucky, USA).

current country characteristics and historical information on selected economic variables for the three countries.

EXHIBIT C Select Country Factors and Historical Economic Data, 2000–2010

	Country Factors	Growth in Hours Worked (%)	Growth in Labor Productivity (%)	Growth in TFP (%)	Growth in GDP (%)
Country A	• High level of savings and investment • Highly educated workforce • Low tariffs on foreign imports • Limited natural resources	0.9	2.4	0.6	3.3
Country B	• Developed financial markets • Moderate levels of disposable income • Significant foreign direct and indirect investments • Significant natural resources	−0.3	1.6	0.8	1.3
Country C	• Politically unstable • Limited property rights • Poor public education and health • Significant natural resources	1.8	0.8	−0.3	2.6

Klymchuk instructs an associate economist at EC to assist him in forecasting the change in stock market value for each country. Klymchuk reminds the associate:

> ***Statement 1:*** "Over short time horizons, percentage changes in GDP, the ratio of earnings to GDP, and the price-to-earnings ratio are important factors for describing the relationship between economic growth and stock prices. However, I am interested in a long-term stock market forecast."

A client is considering investing in the sovereign debt of Country A and Country B and asks Klymchuk his opinion of each country's credit risk. Klymchuk tells the client:

> ***Statement 2:*** "Over the next 10 years, I forecast higher potential GDP growth for Country A and lower potential GDP growth for Country B. The capital per worker is similar and very high for both countries, but per capita output is greater for Country A."

The client tells Klymchuk that Country A will offer 50-year bonds and that he believes the bonds could be a good long-term investment given the higher potential GDP growth. Klymchuk responds to the client by saying:

> ***Statement 3:*** "After the next 10 years, I think the sustainable rate of economic growth for Country A will be affected by a growing share of its population over the age of 65, a declining percentage under age 16, and minimal immigration."

The client is surprised to learn that Country C, a wealthy, oil-rich country with significant reserves, is experiencing sluggish economic growth and asks Klymchuk for an explanation. Klymchuk responds by stating:

> ***Statement 4:*** "While countries with access to natural resources are often wealthier, the relationship between resource abundance and economic growth is not clear. My analysis shows that the presence of a dominant natural resource (oil) in Country C is constraining growth. Interestingly, Country A has few natural resources, but is experiencing a strong rate of increase in per capita GDP growth."

Klymchuk knows that growth in per capita income cannot be sustained by pure capital deepening. He asks the associate economist to determine how important capital deepening is as a source of economic growth for each country. Klymchuk instructs the associate to use the data provided in Exhibit C.

Klymchuk and his associate debate the concept of convergence. The associate economist believes that developing countries, irrespective of their particular characteristics, will eventually equal developed countries in per capita output. Klymchuk responds:

> ***Statement 5:*** "Poor countries will converge to the income levels of the richest countries only if they make appropriate institutional changes."

7. Based on the country factors provided in Exhibit C, the country *most likely* to be considered a developing country is:
   A. Country A.
   B. Country B.
   C. Country C.

8. Based on Exhibit C, capital deepening as a source of growth was *most* important for:
   A. Country A.
   B. Country B.
   C. Country C.

9. Based on Klymchuk's Statement 1, over the requested forecast horizon, the factor that will *most likely* drive stock market performance is the percentage change in:
   A. GDP.
   B. the earnings-to-GDP ratio.
   C. the price-to-earnings ratio.

10. Based solely on the predictions in Statement 2, over the next decade Country B's sovereign credit risk will *most likely*:
    A. increase.
    B. decrease.
    C. not change.

11. Based on Statement 2, the difference in per capita output between Country A and Country B is *most likely* due to differences in:
    A. capital deepening.
    B. capital per worker.
    C. total factor productivity.

12. Based on Statement 3, after the next 10 years the growth rate of potential GDP for Country A will *most likely* be:
    A. lower.
    B. higher.
    C. unchanged.

13. Based on Statement 4 and Exhibit C, the sluggish economic growth in Country C is *least likely* to be explained by:
    A. limited labor force growth.
    B. export-driven currency appreciation.
    C. poorly developed economic institutions.

14. Based on Statement 4, the higher rate of per capita income growth in Country A is *least likely* explained by:
    A. the rate of investment.
    B. the growth of its population.
    C. the application of information technology.

15. The type of convergence described by Klymchuk in Statement 5 is *best* described as:
    A. club convergence.
    B. absolute convergence.
    C. conditional convergence.

The following information relates to Questions 16 through 21.[27]

At a recent international finance and economics conference in Bamako, Mali, Jose Amaral of Brazil and Lucinda Mantri of India are discussing how to spur their countries' economic growth. Amaral believes that growth can be bolstered by removing institutional impediments, and suggests several possibilities for Brazil: launching a rural literacy program, clarifying property rights laws, and implementing a new dividend tax on foreign investors.

Mantri responds that, for India, capital deepening will be more effective, and has proposed the following ideas: building a group of auto and textile factories in the southern

[27]These practice problems were developed by E. Shepard Farrar, CFA (Washington, D.C., USA).

states, developing a north-south and east-west highway network, and sponsoring a patent initiative.

In response, Amaral says to Mantri:

"Based on endogenous growth theory, one of those proposals is more likely to raise total factor productivity than result in pure capital deepening."

While Mantri recognizes that India lacks the significant natural resources of Brazil, she states that India can overcome this challenge by bolstering long-term growth through three channels:

Channel 1: Deepening the capital base.
Channel 2: Making investments in technology.
Channel 3: Maintaining a low rupee exchange rate.

Each country's basic economic statistics were presented at the conference. Selected data for Brazil and India are shown in Exhibit D. Adama Kanté, a fund manager based in Mali, is planning to increase the fund's allocation to international equities, and, after some preliminary analysis, has determined that the new allocation will be to Brazilian or Indian equities. After reviewing the data in Exhibit D, Kanté decides that the allocation will be to Indian equities.

EXHIBIT D Economic Statistics, Brazil and India

Economic Statistic	Brazil	India
GDP per capita, 2010	$9,589	$3,575
GDP per capita growth, 1990–2010	1.62%	4.84%
GDP growth, 2005–2008	4.9%	8.2%
Growth due to labor productivity component	2.9%	6.0%
Growth due to capital deepening component	3.4%	3.6%

Kanté is concerned about the low standard of living in Mali. To improve per capita GDP, Kanté is considering five specific strategies:

Strategy 1: Lower the country's tax rate.
Strategy 2: Introduce policies that encourage the return of highly educated Malian emigrants.
Strategy 3: Build day care centers to permit greater participation of women in the workforce.
Strategy 4: Impose high tariffs on imports to protect the country's nascent industries.
Strategy 5: Use economic development bank loans to improve the country's transport and manufacturing infrastructure.

16. Which of Amaral's initiatives is *least likely* to achieve his stated growth objective?
A. Dividend tax
B. Rural literacy
C. Property rights

17. Which proposal for India is Amaral *most likely* referring to in his response to Mantri?
    A. Patent initiative
    B. Highway network
    C. Auto and textile factories

18. The channel that is *least likely* to help India overcome its challenge of lacking significant natural resources is:
    A. Channel 1.
    B. Channel 2.
    C. Channel 3.

19. Based on Exhibit D, which Indian economic statistic is *least likely* to support Kanté's international equity allocation preference?
    A. GDP per capita
    B. Growth due to labor productivity
    C. Growth due to capital deepening

20. The strategy that is *least likely* to improve per capita GDP in Mali is:
    A. Strategy 1.
    B. Strategy 2.
    C. Strategy 3.

21. Which of the following strategies being considered by Kanté is *most likely* to undermine or delay convergence with developed economies?
    A. Strategy 2
    B. Strategy 4
    C. Strategy 5

CHAPTER 12

ECONOMICS OF REGULATION

Chester S. Spatt

LEARNING OUTCOMES

After completing this chapter, you will be able to do the following:

- Describe classifications of regulations and regulators.
- Describe uses of self-regulation in financial markets.
- Describe the economic rationale for regulatory intervention.
- Describe regulatory interdependencies and their effects.
- Describe tools of regulatory intervention in markets.
- Explain purposes in regulating commerce and financial markets.
- Describe anticompetitive behaviors targeted by antitrust laws globally, and evaluate the antitrust risk associated with a given business strategy.
- Describe benefits and costs of regulation.
- Evaluate effects on an industry, company, or security of a specific regulation.

1. INTRODUCTION

Regulation is an extremely important topic because regulation not only has potential effects at the macro level on the economy, but also has potential effects at the micro level on companies and individuals. Regulation may develop either proactively in anticipation of consequences of changes in the environment or reactively in response to some occurrence(s). For example, changes that resulted from technological advances in the marketplace because of new means of communications and applications of computers have led to a variety of regulation, proactive and reactive. Regulation has also developed in response to financial crises and undesirable behaviors or actions that have occurred in the past.[1] Regulations are necessary because market solutions are not adequate for all situations.

[1]In some cases, these behaviors or actions were criminal and violated existing laws. One goal of regulators is to try to detect these activities earlier.

A significant challenge on the financial regulation front is how to deal with systemic risk (the risk of failure of the financial system) and the consequences of risk taking by financial institutions. On other fronts, such issues as labor regulation, environmental regulation, and electronic privacy are receiving increased attention. Changes in regulatory structure and regulatory uncertainty can have substantial effects on business decisions. One of the significant challenges facing professionals in the finance industry is to anticipate and understand the consequences of potential changes in the regulatory environment and to specific regulations.

Section 2 of this chapter provides an overview of regulation, including classifications of regulations and regulators, roles of regulations, and regulatory tools. Section 3 describes regulation of commerce and areas of focus in commercial regulation. Section 4 describes regulation of financial markets, including securities regulation and regulation of financial institutions. Section 5 describes the assessment of costs and benefits of regulation. Section 6 describes and illustrates an analysis of regulation. Section 7 summarizes the key points of the chapter, and practice problems conclude the chapter.

2. OVERVIEW OF REGULATION

Regulatory frameworks, among other effects, influence how businesses operate. A regulatory framework develops a set of rules or standards of conduct. Regulations may impose restrictions on and/or mandate how businesses interact with others, including other businesses, consumers, workers, and society in general. The regulations may also impose constraints on and/or mandate how businesses operate internally. It is important for an analyst to understand how regulation may affect the business environment, as well as individual industries or businesses. There is a separate discussion of regulation of financial markets, although that is part of the broad business environment. This section includes an overview of regulation, such as the classification of regulations and regulators, roles of regulations, and regulatory tools.

2.1. Classification of Regulations and Regulators

Regulations are sometimes enacted by legislative bodies (often these regulations are laws), but more typically arise from the determination of regulatory bodies. Regulatory bodies may be either governmental agencies or independent regulators (other regulators granted authority by a government or governmental agency). Regulatory bodies have legal authority to enact and enforce regulation within the parameters of the mandate given to them. In many instances, a legislative body enacts a statute at a broad level, leaving it to regulatory bodies to fill in implementation details.[2] Courts play a role in regulation as well—helping to interpret regulations and laws, defining permitted and not permitted regulatory practices, and in some instances, imposing sanctions for regulatory violations. Regulations can be classified as reflecting laws enacted by

[2]This description by the U.S. Securities and Exchange Commission (SEC) is illustrative of how the process works: "Rulemaking is the process by which federal agencies implement legislation passed by Congress and signed into law by the President. Major pieces of legislation, such as the Securities Act of 1933, the Securities Exchange Act of 1934, the Investment Company Act of 1940, and the Sarbanes-Oxley Act, provide the framework for the SEC's oversight of the securities markets. These statutes are drafted broadly, establishing basic principles and objectives. To ensure that the intent of Congress is carried out in specific circumstances—and as the securities markets evolve technologically, expand in size, and offer new products and services—the SEC engages in rulemaking." www.sec.gov/about/whatwedo.shtml

legislative bodies (**statutes**), rules issued by government agencies or other regulators (**administrative regulations** or **administrative law**), and interpretations of courts (**judicial law**).

Although government agencies make many regulatory determinations, **independent regulators** make some regulations. The authority of independent regulators comes from their recognition by a government body or agency, but they are not government agencies per se. One distinction between government agencies and independent regulators is that the latter typically do not rely on government funding. Some argue that an advantage of independent regulators is that they are to some extent immune from political influence and pressure. Some independent regulators are **self-regulating organizations**, private, nongovernmental organizations that both represent and regulate their members. While these organizations may be independent of the government and to an extent immune from political pressure, they may be subject to pressure from their members. Self-regulating organizations given recognition and authority, including enforcement power, by a government body or agency, are independent regulators. However, not all self-regulating organizations are independent regulators. Some self-regulating organizations receive authority from their members, who agree to comply with the organization's rules and standards and its enforcement of these. This authority does not have the force of law; self-regulating organizations are not regulators unless they are given recognition and authority, including enforcement power, by a government body or agency.

In addition, regulatory authorities may reference the work of outside bodies in their regulations. Examples of these outside bodies are accounting standard–setting bodies, such as the International Accounting Standards Board (IASB) and the U.S. Financial Accounting Standards Board (FASB), and credit reporting agencies. Regulatory authorities retain the legal authority to enforce any regulation that references the work of these bodies. In the case of accounting standard–setting bodies—which are typically private-sector, nonprofit, self-regulated organizations—the requirement to prepare financial reports in accordance with specified accounting standards is the responsibility of regulatory authorities. The standard-setting bodies may set the standards, but the regulatory authorities recognize and enforce the standards. Ratings by credit-rating agencies—that are typically private-sector, profit-oriented entities—were often referenced in regulations related to acceptable holding by certain entities. Issues with conflicts of interest, however, have resulted in efforts to reduce references to credit-rating agencies in regulations.[3]

The relatively simple classification of regulators (legislative bodies, government agencies, independent regulators, courts) and regulations (statutes, administrative regulations, and judicial law) is useful, but does not reflect the complexities and nuances that exist with respect to regulators and regulation. In some cases, the classification of a regulator is clear, and in other cases, the classification is ambiguous.

For example, the U.S. Securities and Exchange Commission (SEC), the government agency that regulates the securities markets in the United States, allocates some regulatory responsibilities to specified self-regulatory organizations (SROs). In this context, an SRO is a self-regulating organization and an independent regulator recognized and granted authority by a government agency or body. These SROs are funded independently rather than from the government. The Financial Industry Regulatory Authority (FINRA), one such SRO, describes itself as "the largest independent regulator for all securities firms doing business in the United

[3]Credit-rating agencies often are compensated by the entity requesting the rating; this practice has resulted in questions about the independence and reliability of such ratings. Perhaps the most significant source of conflict of interest is that the issuer selects the ratings that will be purchased and published (ratings shopping).

States. FINRA's mission is to protect America's investors by making sure the securities industry operates fairly and honestly."[4] FINRA has the authority to enforce industry rules and federal securities laws. In this case, it is clear that FINRA is an independent regulator and an SRO. The U.S. Congress established the Public Company Accounting Oversight Board (PCAOB), a nonprofit corporation, to oversee the audits of public companies. Previously, the audit profession was self-regulated. The PCAOB is funded primarily through the assessment of annual fees on public companies, brokers, and dealers. The SEC oversees the PCAOB. The PCAOB is an independent regulator rather than a government agency, but it is not an SRO.

The role of SROs varies among countries. In some countries, such as in the United States, SROs have specific regulatory authority, and in other countries, self-regulating organizations are rarely or never recognized as independent regulators. For example: "One of the many significant recent legislative amendments that was introduced in Australia with the Financial Services Reform Act 2001 was the removal of the official regulatory standing of self-regulatory organizations (SROs). SROs, whether they are exchanges, industry associations, or some other form of peer group, have traditionally set standards of behaviour or codes of conduct for market participants."[5] According to Carson (2011), the role of self-regulation in Europe, with the exception of the United Kingdom, was limited because of civil law systems and resulting reliance on government supervision. In the United Kingdom and other countries with common-law systems, reliance on self-regulation has been more extensive. The roles of SROs in regulation range from nonexistent to having some regulatory authority. Regulators are concerned with the corporate governance of SROs and the management of their conflicts of interest. The extent of the concern is a factor in deciding the regulatory role, if any, of the SRO in question.

In Singapore, "Statutory boards are entities separate from the government, with specific legislation governing their operations. Most, if not all, statutory boards impose charges on some or all of their services. Statutory boards that do not generate sufficient revenue to meet their expenses would receive grants from the government to finance their operations. These grants are funded from the government's annual budget."[6] The statutory boards are described as separate from the government, yet they are subject to specific legislation governing their operations and they may receive government funding. It is ambiguous whether Singapore's statutory boards are government agencies or independent regulators. The Singapore Economic Development Board (EDB), one such statutory board, describes itself as "the lead government agency for planning and executing strategies to enhance Singapore's position as a global business centre."[7] Another statutory board, the Accounting and Corporate Regulatory Authority (ACRA), describes itself as "the national regulator of business entities and public accountants in Singapore."[8] Although the EDB clearly identifies itself as a government agency, it is less clear whether the ACRA, given the description of a statutory board and the description of itself, should be classified as a government agency or as an independent regulator.

Classifying regulatory bodies that exist in unions, such as the Union of South American Nations (UNASUR) and the European Union (EU), can present challenges. For example, the European Commission (EC), which has a mission to promote the general interest of the EU, can issue regulations, directives, and decisions. These are jointly referred to as EU law.

[4]www.finra.org/AboutFINRA/.

[5]www.asic.gov.au/asic/pdflib.nsf/lookupbyfilename/integration-financial-regulatory-authorities.pdf/$file/integration-financial-regulatory-authorities.pdf.

[6]www.ifaq.gov.sg.

[7]www.edb.gov.sg/content/edb/en/about-edb/our-strategy/what-we-do.html.

[8]www.acra.gov.sg/About_ACRA/About_Us.htm.

Regulations have binding legal force throughout every EU member state on a par with national laws.[9] Directives identify desired results and require national authorities to put laws in place to achieve these.[10] Decisions are binding laws addressed to specific parties and are the result of specific cases.[11] Regulations appear to have the characteristics of administrative regulations. Directives appear to have the characteristics of statutes; these are at a broad level and another body needs to fill in the implementation details. Decisions appear more similar to judicial law. Thus, it is hard to classify the European Commission based on the type of regulation issued. In choosing whether to issue a directive or a regulation, the EC appears to take into account the desired outcome. For example, the European Market Infrastructure Regulation (EMIR) takes the form of a regulation rather than a directive. This choice reflects the EC's desire to build a harmonized regulatory framework across the EU. Regardless of how a regulatory body is classified, it is important to identify the regulators and regulations that might affect the entities or industry being analyzed.

Regulations address a broad range of issues and can be classified by their objectives. These include safety (for example, food, products); privacy (for example, financial information); protection (for example, intellectual property); environmental issues (for example, pollution); labor or employment (for example, workers' rights, employment practices); commerce or trade (for example, consumers' rights and protection, investors' protection, antitrust); and the financial system (for example, prudential supervision of institutions, capital requirements, insider trading). It is difficult, if not impossible, to think of an area of life unaffected by regulation.

Although much of the focus of this chapter is on the rules themselves and their development, impact, and implementation, regulatory enforcement and sanctions also play an important role. This division between development and enforcement of regulation also represents a possible way to classify laws or regulation. **Substantive law** focuses on the rights and responsibilities of entities and relationships among entities, and **procedural law** focuses on the protection and enforcement of the substantive laws. Regulators typically have responsibility for both substantive and procedural aspects of their regulations. In developing regulations, the regulator must implicitly consider the roles of regulation.

2.2. Economic Rationale for Regulation

Regulations are necessary because market solutions are not adequate for all situations. Conceptually, this need can be understood best using ideas from economic theory. One of the basic principles in economics is the *fundamental theorem of welfare economics*. Assuming constant returns to scale, no frictions,[12] and no externalities,[13] competitive market (equilibrium)

[9] http://ec.europa.eu/eu_law/introduction/what_regulation_en.htm.

[10] http://ec.europa.eu/eu_law/introduction/what_directive_en.htm. Christensen, Hail, and Leuz (2011) use differences in the timing of the implementation of EU directives tightening the regulation of market abuse and transparency across countries to assess the impact of tighter securities regulation.

[11] http://ec.europa.eu/eu_law/introduction/what_decision_en.htm.

[12] Examples of frictions are costs for or restraints on trading and asymmetrical information.

[13] Externalities are spillover effects of production and consumption activities onto others who did not consent to participate in the activity. A positive externality provides a spillover benefit and a negative externality generates a spillover cost. For example, if one person does home improvements, neighbors may benefit from increases in their home values even though they have expended no resources to improve their properties. Similarly, if one home in a neighborhood is not maintained, neighbors may bear such costs as losses in property values or expenses incurred to keep pests off their property.

allocations[14] are efficient or *Pareto optimal*. There is no way to redistribute resources and make some agents better off without making others worse off.[15] Furthermore, any efficient allocation of resources can be sustained as a market equilibrium for an appropriate set of prices. Hence, absent frictions and externalities, the market solution will be economically efficient, so there would be no benefit to regulatory intervention.

The case for regulatory intervention rests on the presence of **informational frictions** and **externalities**. Informational frictions result in a variety of issues, which regulators attempt to address. These issues include *adverse selection* (private information in the hands of some, but not all, market participants, which affects the consumption of goods or services), and *moral hazard* (incentive conflicts that arise from the delegation of decision making to agents or from contracts that will affect the behavior of one party to the detriment of the other party to the contract). Asymmetrical information, in general, may allow one entity to have an inherent advantage over another entity with which interaction occurs, and the resulting regulation focuses on establishing rights and responsibilities of entities and adjusting relationships among entities.

Many aspects of regulation reflect the provision of public goods, an externality issue. Although the consumption of a private good by a consumer would deny access to the same units by other consumers, many individuals can consume a public good simultaneously, and indeed, it can be difficult to exclude others from the public good. Many regulations represent an attempt to respond to a public goods or externality problem. In effect, because there are shared benefits from consuming the public good, markets would not produce the optimal amount of these goods. Classic examples of public goods include national defense and standard setting.

Some public goods are considered local public goods because those living in particular areas largely reap the benefits. This type of good is relevant to regulation in a global economy because those with jurisdiction undertake many regulatory decisions over particular geographic areas. Indeed, there can be strong spillover effects associated with such goods and even their regulation. Policy makers (regulators and legislators, for example) in various jurisdictions have become sensitive to the spillover effects of their actions.

2.2.1. Regulatory Interdependencies

An interesting facet of regulation is how regulated entities view the regulation. The answer is often context specific—while there are obviously many examples in which regulated companies fight against new proposed regulations, it is far from universal. Regulated company efforts to fight particular regulations tend to attract more public attention than when the companies are sympathetic to the contemplated regulations. Even more fundamentally, academics have argued that regulation often arises to enhance the interests of the regulated (see Stigler 1971), often called the **regulatory capture** theory. For example, regulatory actions and determinations can restrict potential competition and coordinate the choices of rivals. In the interaction between regulated entities and their regulators, the regulated entities may possess considerable expertise and knowledge, and some of the individual regulators may even have had their intellectual roots in the industry or aspire to be in the industry in which the

[14]Market (equilibrium) allocations are ones in which (1) agents maximize utility given relative prices and (2) markets clear.

[15]If resources can be redistributed such that any one agent can be made better off without making any other agent worse off, then the original allocation would not have been Pareto optimal.

regulated entities operate. The interactions between regulated entities and their regulators may reinforce the perception (or reality) of regulatory capture.

Regulatory differences across jurisdictions can lead to shifts in location and behavior of entities because of **regulatory competition** and **regulatory arbitrage**. Regulators may compete to provide a regulatory environment designed to attract certain entities (regulatory competition). Entities may engage in regulatory arbitrage; for example, they may identify and use some aspect of regulations that allows them to exploit differences in economic substance and regulatory interpretation or in foreign and domestic regulatory regimes to their (the entities') benefit.

Interdependence in the actions of regulators with different objectives is important in the international arena. Many regulatory issues are relatively common ones around the globe. The commonality reflects both similarities in the challenges confronting different countries and the diffusion of the underlying problems around the globe. While issues such as systemic risk, moral hazard, global warming, and nuclear power regulation all reflect significant global concerns, regulators in different jurisdictions can have different perspectives or face different trade-offs when addressing specific issues. These different perspectives can lead to differences in regulatory treatments. Although such differences are often well justified, regulatory competition can undercut the effectiveness of enhanced regulation in particular countries and impose constraints on regulation. Sometimes regulatory cooperation and coordination are called for in an increasingly interconnected global economy.

An example in the aftermath of the global financial crisis, which was first identified as such in 2008 (hereafter referred to as the 2008 global financial crisis although it lasted beyond 2008), concerned the push toward centralized clearing rather than bilateral settlement of derivatives transactions. Many European and Asian regulators were slower to respond than U.S. regulators. In the United States, the Dodd-Frank Act called for derivatives reforms to be fully resolved by July 2011, whereas the G-20 called for action by member nations by the end of 2012. A number of U.S.-based entities expressed fear that the U.S. markets would be greatly disadvantaged because of the extent of differences in the ultimate regulatory regimes.

In another example, consider issues related to global warming and pollution. How should governments manage and coordinate efforts and adjustments around the globe? The relevant externality is not simply within countries, but across countries. One of the challenging aspects of this issue is that countries are in very different situations. What are the institutional and governance mechanisms that would be appropriate to address this issue on a global basis? Although an economist's solution to the problem of pollution externalities might be to tax it in order to allocate the pollution to the parties that can absorb the cost, that leaves many questions open. How should one allocate permits to pollute among countries? Should countries have the right to pollute related to their past pollution? If not, how would one accommodate differences in living standards? How should one address the equity issues associated with low-wealth and developing countries having a potential comparative advantage in absorbing pollution?[16]

The point of this overall discussion of interdependencies across jurisdictions is not to suggest global governance or a global regulator, but to recognize the reality and implications of diverse trade-offs and preferences among regional, national, and local regulators. To a degree, the presence of diverse and arguably competing jurisdictions influences the stances of national

[16]A memo that Larry Summers wrote in 1991 while chief economist of the World Bank suggested that poor countries should bear much of the pollution (with compensation); the memo resulted in a political firestorm.

and regional regulators. Evidence exists that governments recognize the necessity for global regulatory cooperation and coordination. For example, the Basel Accords establish and promote internationally consistent capital requirements and risk management practices for larger international banks. The Basel Committee on Banking Supervision, among other functions, has evolved into a standard setter for bank supervision. The International Organization of Securities Commissions (IOSCO) is a self-regulating organization but not a regulatory authority. Its members regulate a significant portion of the world's capital markets. This organization has established objectives and principles to guide securities and capital market regulation, and its members agree to adhere to these.[17]

At the country level, the objectives of diverse government regulators can differ and potentially lead to regulations that seem inconsistent. Bank supervisors (whether as a function of the central bank, another entity, or combination of entities) focus on **prudential supervision**—regulation and monitoring of the safety and soundness of financial institutions in order to promote financial stability, reduce systemwide risks, and protect customers of financial institutions. The objectives of securities commissions, per IOSCO, are protecting investors; ensuring that markets are fair, efficient, and transparent; and reducing systemic risk. In some situations, these goals are quite different in their implications. The bank supervisor may be reluctant or even unwilling to release the results of the bank's tests of financial institutions in order to promote financial stability and to avoid systemic risk because of a loss of confidence. The securities commission is more likely to advocate for the release of information that might be relevant to investors (see Spatt 2009).

A general conclusion is that regulation by different regulators, even with seemingly similar objectives, can lead to very different regulatory outcomes. The causes of this variation include different orientations of the regulators and objectives that are broadly stated or ill defined.

2.3. Regulatory Tools

Given a range of regulatory tools and measures, it is important to recognize that regulatory and governmental policies should be predictable as well as effective in achieving objectives. It is very difficult for any entity to function with confidence and success in an environment where the rules are unclear or in a state of flux (in other words, where there is considerable regulatory uncertainty). Regulatory choices or government policies that will be consistent over time are desirable. The most effective way to ensure time consistency is to focus on regulatory choices that the government will have an incentive to carry out over time. If these choices occur, the regulatory environment is likely to be stable despite the fact that, in many countries,

[17]The member agencies currently assembled in the IOSCO have resolved, through its permanent structures:

- to cooperate in developing, implementing, and promoting adherence to internationally recognized and consistent standards of regulation, oversight, and enforcement in order to protect investors, maintain fair, efficient and transparent markets, and seek to address systemic risks;
- to enhance investor protection and promote investor confidence in the integrity of securities markets, through strengthened information exchange and cooperation in enforcement against misconduct and in supervision of markets and market intermediaries; and
- to exchange information at both global and regional levels on their respective experiences in order to assist the development of markets, strengthen market infrastructure, and implement appropriate regulation. www.iosco.org/about/.

governmental decision makers (with diverse political preferences) change on a regular basis. It is helpful to utilize regulatory tools that are consistent with maintaining a stable regulatory environment. Regulatory tools and government interventions in markets include the use of price mechanisms, such as taxes and subsidies; regulatory mandates and restrictions on behaviors, including establishing rights and responsibilities; provision of public goods; and public financing of private projects.

The issue of how to address pollution is a classic example in regulation. By taxing polluters (or subsidizing those who do not pollute by using a suitable baseline), one can create a system in which marginal incentives are equated across economic agents. The advantage of such an arrangement is that, theoretically, the rights to pollute are redistributed in an efficient manner relative to a fixed allocation. In particular, the structure of the regulation allows market incentives to redistribute the pollution rights to those for whom they are the most valuable at the margin. There are important issues, however, about how to initially establish and distribute the amount of acceptable total pollution. In some situations, historical usage (amount of pollution produced) is used to allocate acceptable levels. One problem is that marginal incentives may be altered in anticipation of this allocation. In other situations, the allocation is the outcome of a political process, which can lead to considerable lobbying. At the heart of this example is the use of a price mechanism to create the appropriate marginal incentives and an efficient allocation of resources. The Coase theorem states that if an externality can be traded and there are no transaction costs, then the allocation of property rights will be efficient and the resource allocation will not depend on the initial assignment of property rights.

Governments can intervene in markets in ways other than through the price mechanism. These include restricting some activities (for example, insider trading, short selling); mandating some activities (for example, capital requirements for banks, registration with a securities commission for certain activities); providing public goods (for example, national defense, transportation infrastructure); and financing private projects (for example, loans to individuals or companies for specified activities that the government deems desirable to encourage). The extent of government provision of public goods and government financing of private projects depends on a number of factors, including the political philosophy of the country or government in power, the structure of the government, and the country's gross domestic product. The problem of **systemic risk** (the risk of failure of the financial system) as a result of the failure of a major financial institution has emerged as an issue in many countries around the world in the aftermath of the 2008 global financial crisis. Systemic risk and **financial contagion** (a situation in which financial shocks spread from their place of origin to other locales; in essence, a faltering economy infects other, healthier economies) are examples of negative externalities. In the EU, the European Systemic Risk Board (ESRB), formed in December 2010, is an independent EU body tasked with macro-prudential oversight of the EU financial system. The Dodd–Frank Act enacted by the U.S. legislative body attempts to mitigate systemic risk, among other objectives. U.S. regulatory bodies (rather than the legislature) are largely responsible for implementing the provisions of the Dodd–Frank Act. At the same time, policy makers may not have sufficiently clarified how to evaluate differences in systemic risk imposed by different financial institutions or even defined systemic risk except in broad terms.

It is difficult to assess the extent to which the new approaches created by legislation, such as the Dodd–Frank Act, and other regulatory changes, such as the creation of the ESRB, will reduce systemic risk. There are a number of reasons for this difficulty. The amount of underlying empirical data about systemic crises is very limited. By definition, these events are outliers on some metrics, so the types and sources of future crises are likely to be rather

different. Regulations designed with a prior crisis in mind may not head off a future crisis (or even contain the seeds of one). It can be difficult to assess the potential effectiveness of regulatory actions before an event and even after the fact. The mere fact that a crisis does not occur is not necessarily evidence that the regulation was the reason that a crisis did not occur. It is also plausible that some regulatory responses have the unintended consequence of mitigating one source of risk while increasing another source of risk. All of these issues make effective regulation challenging to design.

Generally, more than one regulatory approach is feasible and worthy of consideration in a specific situation. Two examples that illustrate a range of possible regulatory responses are (1) conflict of interest policies and (2) trading restrictions on insiders. To illustrate the first situation, consider a hypothetical situation in which a potential employee of a regulator has some degree of financial exposure to a regulated company. Such exposure could come about in many ways (for example, spousal employment, a marketable position in an investment portfolio, an illiquid position resulting from past employment) and at a variety of financial levels. What types of regulatory restrictions might arise? Among the potential regulatory responses are the following: The individual is barred from employment at the agency. The individual is barred from working on specific (or all) projects involving the company in question. The individual sells the position; the sale can be voluntary or mandated. The individual is required to disclose the nature of the potential conflict to higher-level decision makers to whom the individual will be providing recommendations. Broadly, the alternative remedies include a bar on involvement, resolution of the conflict, or disclosure of it.

Turning to the case of corporate insiders, there are both potential regulatory and corporate restrictions. Examples of regulatory responses are a ban from trading on nonpublic information and a requirement that insiders disclose trades. The company may impose a blackout period during which insiders are banned from trading on the company's stock (these periods often precede earnings announcements and continue shortly afterward). The appropriate remedy is dependent on the underlying facts and circumstances, and arguably the appropriate standards would reflect the specific context. As stated previously, there often are alternative ways to tackle a particular regulatory issue. When evaluating potential regulatory responses to an issue and the effects of the potential regulation, it is important to consider a range of feasible responses.

An important aspect of effective regulation is the potential ability to impose sanctions on violators of the regulations; in other words, it is important to be able to enforce the regulations. IOSCO clearly identifies this aspect as one of the agreed-upon principles of securities regulation: "The Regulator should have comprehensive enforcement powers."[18] Enforcement of securities regulations and regulations on businesses may include sanctions on the violating corporation (business or company), the individual violator(s), or both. Corporate sanctions may be appropriate when the company caused harm to others. The sanctions often involve monetary fines, fees, or settlement, and in the case of individuals, the sanctions may involve prison terms. However, in some situations, such as cases of accounting fraud, stockholders may actually be the victims. In such instances, when the stockholders were harmed by the wrongdoing, the case for sanctions, such as fines, against the company is far from compelling. The sanctions may simply redistribute funds from current stockholders to the stockholders who were the specific victims, and the company incurs real resource costs.

[18]International Organization of Securities Commissions, "Objectives and Principles of Securities Regulation," June 2010.

For various reasons, it can be difficult to prosecute or achieve settlements with individual violators. First, it often is difficult to detect violations and to identify exactly which individuals were at fault. Furthermore, the individuals possess very strong incentives to fight in order to protect their reputations and livelihoods. Indeed, individuals are often able to fight using corporate resources because of indemnification provisions in their employment contracts. The intent of these provisions may be to protect risk-averse executives against inadvertent liability and to potentially align their interests with the stockholders' interests, but the provisions may result in protecting an executive to the detriment of the stockholders. The incentive to fight individual sanctions may be especially strong because of not only financial costs, but also other costs, such as reputational costs.

EXAMPLE 12-1 Overview of Regulation

Lee Ming, an analyst, is researching the use of self-regulation in securities markets by reading "Self-Regulation in Securities Markets" by John Carson (2011). The main factors identified as contributing to a trend of decreased reliance on self-regulation in securities markets are:

- Privatization of securities exchanges.
- Intense competition.
- Uncertainty regarding the effectiveness of self-regulation.
- Internationalization.
- Strengthening of government regulators.
- Trend toward consolidation of financial regulators.
- Cooperative regulation.
- Pressure to increase efficiency and lower costs.

Reasons there still is reliance on self-regulation include:

- It increases the overall level of regulatory resources.
- It uses the knowledge and expertise of industry professionals.
- It enables the regulator to focus on other priorities while relying on an SRO for frontline supervision of its members and regulated markets.

In addition, Ming makes note of the following statements, among others, in Carson's paper:

Statement 1: "In much of the world, the value of self-regulation is being debated anew. Forces, such as commercialization of exchanges, development of stronger statutory regulatory authorities, consolidation of financial services industry regulatory bodies, and globalization of capital markets, are affecting the scope and effectiveness of self-regulation." (p. 2)

Statement 2: "IOSCO states that use of an SRO may be appropriate where an SRO exists that has the capacity to carry out the purpose of regulation and to enforce

compliance with rules by its members, and where the SRO is subject to adequate oversight by the regulator." (p. 7)

Statement 3: "France employs a similar approach, because regulations may provide that firms have a duty to implement standards set by associations such as the *Association française des marchés financiers* (AMAFI)." [Note that this organization was previously known as the French Association of Investment Firms.] (p. 8)

Statement 4: "Even in countries where formal SROs [with regulatory powers] do not exist, as in Europe, a trend is observed toward increased use of securities industry bodies to support the regulatory system by providing guidance, codes of conduct, continuing education and so on." (p. 12)

Statement 5: "In many countries, the cost of government 'bailouts' and the need for significant government intervention in the financial system have raised demands for regulatory reform. Potential reforms may take several forms, ranging from (1) adoption of stronger laws and regulations to (2) changes in the structure of regulatory systems (including consolidating financial regulators), (3) improved governance and accountability of financial regulators, and (4) stronger supervision of compliance with laws and rules by financial institutions." (p. 53)

1. In Statement 1, commercialization of exchanges *most likely* led to debates on and decline in reliance on self-regulation because of concerns about:
   A. regulatory capture.
   B. regulatory arbitrage.
   C. regulatory competition.
2. Given an objective of fair and efficient markets, the *most* important criterion from Statement 2 when using an SRO for regulatory purposes is:
   A. capacity of the SRO.
   B. adequate oversight by the regulator.
   C. ability of the SRO to enforce compliance.
3. Considering Statements 3 and 4, the Association française des marchés financiers is *most likely*:
   A. an independent regulator.
   B. a self-regulating organization.
   C. both an independent regulator and a self-regulating organization.
4. In response to the information in Statement 5, governments are *least likely* to increase reliance on:
   A. government agencies.
   B. independent regulators.
   C. self-regulating organizations.
5. Globalization of capital markets is *most likely* to result in increased concerns about:
   A. contagion.
   B. regulatory competition.
   C. both contagion and regulatory competition.
6. The regulatory tools *least likely* to be used by self-regulating organizations are:
   A. price mechanisms.
   B. restrictions on behaviors.
   C. provision of public goods.

Solution to 1: A is correct. Regulatory capture has always been a concern when SROs are used, but commercialization of exchanges has led to increased concern about conflicts of interest and regulatory capture.

Solution to 2: B is correct. Adequate oversight by the regulator is a critical aspect in ensuring that the SRO fulfills its roles in a manner consistent with a fair and efficient market.

Solution to 3: B is correct. The Association française des marchés financiers is a self-regulating organization that issues standards and guidance that are referenced by a regulator, but it is not a regulator, independent or otherwise; it is also not a formal SRO with regulatory powers. It is a standard-setting body but it has no regulatory authority; regulatory authorities recognize and enforce the standards. This role is similar to the role of many accounting standards boards.

Solution to 4: C is correct. Governments are least likely to increase reliance on self-regulating organizations. There appears to be a desire for increased government intervention, stronger laws, and improved governance and accountability. The use of SROs is not consistent with these.

Solution to 5: C is correct. Globalization is likely to result in increased concerns about contagion and regulatory competition. It is easier for a financial shock to spread. A government may use its regulatory environment as a basis to attract entities from around the world.

Solution to 6: A is correct. SROs are least likely to use price mechanisms. They typically regulate behaviors and often provide public goods in the form of standards.

3. REGULATION OF COMMERCE

Given the amount of regulation in existence, it is useful to have a framework within which to consider regulation. IOSCO developed a framework of matters to be addressed in the domestic laws of a jurisdiction to facilitate effective securities legislation. This framework is shown in Exhibit 12-1.

The framework is a useful, but by no means exhaustive, list of areas of regulation relevant to an analyst. For example, labor, consumer protection, and environmental laws, which are not included in the list, may significantly affect a business or industry. These laws often address issues of safety and health. Awareness of the basic types of laws that affect economies, financial systems, industries, and businesses is useful to an analyst. This knowledge will help the analyst to identify areas of concern and to consider proactively potential effects of regulations, existing and anticipated.

As discussed previously, externalities (such as pollution) and public goods problems are critical to the operation of our national and global economies. Similarly, it is difficult to structure private markets for many of the kinds of infrastructure decisions that are central to the operation of society and the economy. The role of government regulation is critical to setting out an underlying framework and facilitating business decisions that involve a considerable degree of coordination.

EXHIBIT 12-1 IOSCO's Legal Framework

Effective securities regulation depends on an appropriate legal framework. The matters to be addressed in the domestic laws of a jurisdiction include:

1. Company Law
   - 1.1 company formation
   - 1.2 duties of directors and officers
   - 1.3 regulation of takeover bids and other transactions intended to effect a change in control
   - 1.4 laws governing the issue and offer for sale of securities
   - 1.5 disclosure of information to security holders to enable informed voting decisions
   - 1.6 disclosure of material shareholdings
2. Commercial Code/Contract Law
   - 2.1 private right of contract
   - 2.2 facilitation of securities lending and hypothecation
   - 2.3 property rights, including rights attaching to securities, and the rules governing the transfer of those rights
3. Taxation Laws
   - 3.1 clarity and consistency, including, but not limited to, the treatment of investments and investment products
4. Bankruptcy and Insolvency Laws
   - 4.1 rights of security holders on winding up
   - 4.2 rights of clients on insolvency of intermediary
   - 4.3 netting
5. Competition Law
   - 5.1 prevention of anti-competitive practices
   - 5.2 prevention of unfair barriers to entry
   - 5.3 prevention of abuse of a market dominant position
6. Banking Law
7. Dispute Resolution System
   - 7.1 a fair and efficient judicial system (including the alternative of arbitration or other alternative dispute resolution mechanisms)
   - 7.2 enforceability of court orders and arbitration awards, including foreign orders and awards

Source: Annexure 3 in "Objectives and Principles of Securities Regulation," May 2003, www.iosco.org/library/pubdocs/pdf/IOSCOPD154.pdf.

The relevant decisions arise at a number of levels. Arguably, many of these would be within the domain of national governments, but some of the relevant externalities are global. While common examples involve environmental issues, such as pollution, global warming externalities across countries, and externalities associated with nuclear waste storage, there are other relevant externalities in a global economy. It is important to have international mechanisms to facilitate the coordination and acceptance of responsibilities across national governments (typically, national governments are best able to coordinate decisions within their respective countries). Some of these externalities have long-term consequences (costs) and implications. In fact, arguably some of these long-run consequences may be ones that are difficult to fully quantify and assess.

The role of governments is crucial for promoting commerce locally, nationally, regionally, and globally. Trade agreements are important to global commerce. Government is in a position to facilitate basic features of the business environment, such as establishing the legal framework for contracting and setting standards. Regulation is central to fundamental aspects of our labor markets, such as workers' and employers' rights and responsibilities and workplace safety. Immigration issues are handled through regulation. Fundamental safety regulations with respect to drugs (including the reliance on testing), food products, medical devices, and pollution are significant.

Several issues have emerged as particularly relevant in the context of globalization and the Internet. One issue is the recognition and protection of intellectual property. Government policies regulate intellectual property, prescribing standards and processes that define and govern patents, trademarks, and copyrights. Although the legal standards are country specific, most countries recognize the importance of protecting intellectual property. At the same time, lack of enforcement and protection of intellectual property has emerged as an issue in some of the trading disputes around the globe. Setting technical standards is another issue, given the focus on technology and electronic tools and resources. Even something as mundane as establishing domain names and the related standard setting requires some appropriate delegation of authority.

Privacy issues also have arisen in the context of the Internet. Privacy is particularly important with respect to medical, financial, academic, and employment records. Entities, including businesses and governments, must be protective of the confidential data in their possession and maintain appropriate security procedures. The Internet raises a broad set of issues involving privacy because of the depth of information potentially available about a person's situation (financial and personal), activities, interactions, and purchases. How Internet software navigates these privacy concerns will influence the perceptions and actions of regulators, as well as the acceptance of software innovations and business models in the marketplace.

An appropriate legal environment is crucial for the successful operation of commerce. Clearly defined rules governing contracts, their interpretation, and each party's legal rights under a contract are necessary. A framework for financial liability and dealing with bankruptcy is necessary for suitable private incentives to enter into economic contracts, particularly those that require long-term commitments. Such activities as construction projects, energy exploration and extraction projects, and even such mundane activities as relocation decisions involve significant long-term, dynamic commitments. Precommitment by society to a well-defined set of rules and standards is crucial to facilitating the willingness of market participants to engage in long-term commitments.

For example, consider the situation in which a company needs to incur significant costs to start a project. These costs are unrecoverable if the project does not progress forward; in other words, these are sunk costs. Without a strong legal framework, the party expending the sunk costs would be reluctant to incur these because of the potential of a holdout problem in which the other side exploits the sunk costs to force renegotiation of the deal. Such contractual difficulties would destabilize the operation of businesses and weaken the economy.

One important role of a national government is to support and protect domestic business interests. A crucial issue in international economic negotiations is protecting businesses against unfair competition. An example of unfair competition is the ability of a company or companies from country X to sell goods at significantly lower prices than its competitors from other countries because of subsidies from government X. Protection of domestic businesses can take the form of tariff and/or nontariff barriers. These protective mechanisms are

sometimes challenged in the international context as giving domestic companies an unfair competitive edge. Analogously, international disputes about whether a country is manipulating or fixing its currency price often center on issues related to competitiveness.

Economics emphasizes the principle of comparative advantage and the value of free trade. Comparative advantage suggests that all countries should allocate their efforts to those goods or services for which development efforts are most productive at the margin. In some cases, protecting or encouraging domestic production, through mechanisms such as subsidies, can impose excess costs on certain sectors or on the broader society. Any potential benefit associated with the subsidy should be compared with the potential distortion created by the government subsidy and the efficiency of the transformation between products. Basic economic principles focus on avoiding the distortion associated with the relative pricing of products because of the underlying transformation process. Similarly, restricting foreign ownership or imposing capital flow restrictions can provide some protection to a domestic economy, but these come at a cost.

Interestingly, whereas in a global context an implicit regulatory goal of government may be to restrict competition from other countries, in a domestic context a regulatory goal is to promote competition (this goal can alternatively be viewed as monitoring and preventing activities that restrict or distort competition). There are several dimensions to this goal. Regulatory approval is typically required for mergers with and acquisitions of major companies. Regulators can effectively block a merger or acquisition, or suggest remedies to resolve a perceived issue (for example, divestiture of particular segments of the businesses to resolve an antitrust issue). When there are competing bids, the actions of the regulator can effectively decide the outcome based on the regulator's assessment of the effects of each bid. Considering the expected response of regulators on competition or antitrust grounds is a central aspect to the evaluation of mergers and acquisitions. An important question to consider, as is done by the regulators, is whether the merger will lead to the monopolization of particular markets and, if so, whether there are ways (such as by divestiture of specific geographic, product, or other segments) to avoid the particular problem.

Competition and antitrust laws also typically prohibit abusive and anticompetitive behavior, such as collusion on prices by companies that dominate a market. Some of the types of behavior that are problematic (beyond mergers that create monopoly power) include exclusive dealings and refusals to deal, pricing discrimination, and engaging in predatory pricing. In response to antitrust issues, regulators not only may impose monetary sanctions but may require companies to alter their business (for example, divest portions or change operating/marketing practices).

The definition of software products and the bundling of them have been an increasing concern with respect to competition. For example, in the late 1990s, Microsoft was subject to a significant challenge by the United States concerning bundling its web browser with the Windows operating system—the crucial issue was whether the bundling reflected innovation or an attempt to monopolize the browser market. Challenges to rivals under competition laws also represent a business strategy. An example of such a challenge is Microsoft's claim in Europe that Google is unfairly impeding competition in the search engine market.

A significant issue that companies need to face in addressing antitrust (or lack of competition) issues is that in many cases they need to satisfy simultaneously a range of regulators. For example, a company may have to satisfy both the U.S. Department of Justice and the European Union if the company plans to use a common product and market strategy across jurisdictions. Despite language and cultural differences, it often will be advantageous to follow a unified strategy around the globe because of business imperatives and likely overlapping

views among regulators of competition. Many of the cases that are significant for the U.S. market are taking place in Europe or elsewhere.

4. REGULATION OF FINANCIAL MARKETS

The regulation of securities markets and financial institutions is especially important because of the consequences to society of failures in the financial system. These consequences range from micro-level to macro-level effects. Potential consequences include financial losses to specific parties, an overall loss of confidence, and disruption of commerce. These consequences were evident in the 2008 global financial crisis. Securities regulation focuses on such goals as protecting investors, creating confidence in markets (a challenging subject), and enhancing capital formation. Although it is difficult to define tangibly what types of regulatory changes would enhance confidence in the financial system, increasing confidence is at least occasionally cited as one of the motives for securities regulation. Many of the rules oriented toward equitable access to information (which, in turn, encourages capital formation) and protecting small investors implicitly serve the role of promoting confidence in the markets.

Regulators of financial institutions focus on such issues as protecting consumers and investors, ensuring safety and soundness of financial institutions, promoting smooth operation of the payments system, and maintaining access to credit. Other macroeconomic concerns of financial regulators include price stability, levels of employment/unemployment, and economic growth.

A key focus of regulators is maintaining the integrity of the markets and acting as a referee for its fairness. This role is distinct compared with financial stability regulation, which is more directly focused on specific outcomes. In addition to securities registration requirements, to facilitate and support the marketplace and the confidence of investors, disclosure requirements are important. Disclosures allow investors to use available information to assess the consequences for investing in and valuing financial instruments and to allow markets to operate. Securities market disclosures occur at various levels, in various forms, and with varied and sometimes unexpected consequences. For example, the Sarbanes-Oxley Act, which required *timeliness* in disclosure of insider transactions,[19] largely resolved the problem of options backdating in the United States—although the architects of that legislation were not aware of the backdating issue.

The disclosure framework is wide-ranging and has high potential importance. The disclosure framework includes financial reporting requirements and accounting standards, prospectus disclosure requirements in conjunction with both securities offerings and annual reports, disclosure requirements in the context of proxy proposals and contests, mutual fund disclosure rules, and price transparency disclosure rules. Disclosure requirements tend to be oriented toward the protection of and provision of information to investors (whether used by investors directly or by service providers).

Many of the regulations governing securities markets are oriented toward mitigating agency problems that arise through delegation to intermediaries. For many financial transactions, parties need to act through others (agents), leading to the potential for agency conflicts. Among the range of examples of regulations addressing potential agency conflicts are

[19]Disclosure of insider transactions has been a long-standing requirement in the securities markets in the United States.

those related to mutual fund fees and governance, the governance of listed companies, rules for proxy voting in companies, best execution requirements on broker/dealers, and treatment of so-called soft dollar expenses by investment advisers in the trading process.

Securities regulators historically have tended to focus more directly on protecting retail investors (individual investors with modest resources and arguably less investment expertise). This tendency has resulted in a lesser focus on financial regulation of hedge funds, private equity, and venture capital funds because of the type of investors (institutional and affluent individual investors) that invest in these funds. For these larger investors, regulators have taken more of a "buyer beware" orientation. For larger investors, it is more difficult to define suitability standards. One approach is to require a more modest range of disclosure requirements related to offering memorandums for a variety of different types of transactions as well as basic antifraud rules. These modest regulations coupled with access requirements that are at least arguably related to the sophistication of investors have typically been the extent of regulation of hedge funds, private equity, and venture capital funds. The majority of securities regulations focus on protecting small investors.

Issues related to prudential supervision of financial institutions and financial stability were introduced earlier. Prudential supervision is regulation and monitoring of the safety and soundness of financial institutions in order to promote financial stability, reduce systemwide risks, and protect customers of financial institutions. This supervision is critical because of the cost that failure of a financial institution can impose on the economy and society. The failure of a bank can result in loss of savings and access to credit. The failure of an insurance company can result in unanticipated losses to those insured. If government-sponsored entities provide protection against these losses or the government chooses to cover all or a portion of these losses, the losses can be spread across a greater sector of society than those directly affected. Additionally, the resulting loss of confidence in the financial system can have far-reaching consequences. The types of regulations include those that focus on diversifying assets, managing and monitoring risk taking, and ensuring adequate capitalization. Monitoring and supervision are important aspects of the regulations. In addition, regulators may set up funds to provide insurance against losses and mandate that premiums or fees be paid into these funds. The benefits of regulation generally do not come without associated costs. For example, regulations that result in the provision of insurance on certain activities may create a moral hazard situation and result in greater risk-taking incentives.

5. COST-BENEFIT ANALYSIS OF REGULATION

In assessing regulation and regulatory outcomes, it is important to assess the overall benefits and costs of regulatory proposals, to develop techniques to enhance the measurement of these, and to examine how economic principles guide regulators.[20] The general benefits of regulation as discussed in earlier sections may be clear, but the measurement of the full impact of the regulation (both benefits and costs) can be challenging. In conducting cost-benefit analysis of regulation, it often is easier to assess the costs of regulation.

Regulatory burden refers to the costs of regulation for the regulated entity; this cost is sometimes viewed as the private costs of regulation or government burden. **Net regulatory**

[20]This theme, and especially the importance of measurement and suitable statistical methods in policy formulation, is examined in more detail in Spatt (2011).

burden is the private costs of regulation less the private benefits of regulation. Understanding the regulatory process will help an analyst recognize the types of challenges that regulators and policy makers face and formulate expectations of regulatory outcomes. Costs and benefits of regulation are important to consider, but often difficult to assess. Many regulators focus narrowly on the implementation costs of regulation (for example, how many compliance attorneys at what cost will need to be hired), but in many instances the most significant costs are the indirect ones that relate to the way in which economic decisions and behavior are altered and market allocations changed.

Regulators view some of the costs associated with regulations as unintended, but it is important to distinguish between two types of such costs. There may be implementation costs that were unanticipated (for example, it turns out more compliance lawyers need to be hired than originally thought) and indirect costs because of unintended consequences. It is important for regulators to recognize that their evaluation of potential rules should reflect the possible unintended consequences as well as the consequences that were the direct object of the rule making. Furthermore, regulatory filings in response to proposed regulations identify at least some of the so-called unintended consequences prior to the implementation of the regulation. It is difficult to argue that such consequences were unanticipated and unintended if they were identified prior to the implementation of the regulation. Unintended consequences are reflective of underlying policy risk and may result in high, unanticipated costs.

Regulatory costs and benefits are especially difficult to assess on a prospective basis compared with a retrospective basis. An after-the-fact analysis allows a comparison of the items of interest before and after the regulation occurs. This comparison allows for a more informed assessment of a regulation because the actual costs and benefits may be identifiable. Even a trial or pilot analysis may be appropriate and helpful in some instances (perhaps too complex to achieve in other instances) to more fully understand the import in advance of a proposed regulation. A potentially feasible and relevant approach in the context of an environment with frequent trading is to use natural experiments and trial phase-ins to generate data suitable for careful cost-benefit analysis.[21] This approach facilitates the assessment of statistical evidence to evaluate the effects prior to the full-blown implementation of the proposed regulation. Such approaches are more feasible for a trading rule in a market with high trading frequency that will generate considerable data and run little risk of disrupting the real economy.

Some regulators undertake relatively little retrospective analysis and assessment of the impact of previously enacted regulations. There has been some call for *sunset provisions* by which a regulation being implemented would be automatically removed after a number of years unless the regulator took further action.[22] The use of sunset provisions would require regulators to undertake a new cost-benefit analysis to continue the regulation. An area of concern is whether regulators devote sufficient attention to assessing the consequences of their past actions. Greater focus on the economic impact of prior decisions would help enhance accountability. A postimplementation review, as is the case with any decision-making process, is a logical step.

Within the United States, administrative law requires that federal regulatory agencies conduct a cost-benefit analysis to assess the consequences of their actions. Court rulings have struck down regulatory actions because inadequate economic and cost-benefit analyses were performed. For example, the U.S. Circuit Court of Appeals overturned the 2004 SEC

[21]Among the contexts in which such techniques have been utilized by U.S. securities regulators have been rules involving short sales, posttrade price reporting, and the tick size increment for trading.
[22]See, for example, Romano (2005).

rule requiring that mutual funds have independent chairs and at least 75 percent independent directors on such grounds.[23] More recently, as reported in the *Economist*, "An appeals court rejected the Securities and Exchange Commission's proxy-access rules. . . . The judges ruled that the regulator had carried out insufficient cost-benefit analysis."[24] In the aftermath of the adoption of the Dodd-Frank Act in the United States, a number of legislators have expressed concern about the quality of the cost-benefit analyses of the financial regulatory agencies responsible for implementing the provisions. This concern reveals an interesting perspective for assessing regulations.

Ideally, regulatory judgments should reflect economic principles and full consideration of the economic costs and benefits rather than the preferences of the current decision makers. Although the potential failure of the fundamental theorem of welfare economics suggests the potential relevance of regulation, it is important to use economic principles to identify and assess alternative remedies and specific actions.

6. ANALYSIS OF REGULATION

The effect of regulations can range from macro-level effects on an economy to micro-level effects on a business. These ultimately have implications for security analysis and valuation. Because regulations are constantly evolving, it is important to monitor issues of concern to regulators and ongoing developments to assess the implications of potential regulation. Using a framework, such as that shown in Exhibit 12-1, can help an analyst identify and focus on particular areas that potentially have significant effects on the industry or entity being analyzed.

When a regulatory environment shifts, lobbying by the potentially affected industry or business may occur; the affected industries and companies are anxious to convey their perspective on the impact of proposed regulations. Potential new regulation may be perceived as either costly or beneficial to the affected entities. Some regulations may work against particular market sectors or industries, while others may work in their favor. If regulators are captive to those that they regulate, regulation is more likely to benefit those regulated. For example, regulation can create demand for particular products and can act as a barrier to entry against rivals. Regulation can change relative demands among products.

One interesting example is the effect of the SEC's Regulation National Market System (NMS) regarding competition among equity trading platforms in the United States. Regulation NMS, an example of a proactive regulation, was intended to take advantage of technological advances to achieve the objectives of efficient, competitive, fair, and orderly markets. Since the 2005 adoption of Regulation NMS, the market share of the trading floor of the New York Stock Exchange (NYSE) has fallen by about two-thirds. Prior to Regulation NMS, NYSE specialists or market makers could take up to 30 seconds to react to orders sent by other platforms. The other platforms were checking whether the NYSE would execute at a more favorable price than the original platform quoted. This process provided considerable opportunity for an NYSE specialist to observe subsequent pricing and to exploit the implicit optionality in the process. This process also made it hard for the rival platform to compete. Consequently, the NYSE could position itself to attract and concentrate much of the market

[23]See *Chamber of Commerce v. SEC*, 412 F.3d 133 (D.C. Cir. 2005) and 443 F.3d 890 (D.C. Cir. 2006).

[24]*Economist*, 30 July 2011, 7. See also *Business Roundtable and Chamber of Commerce v. SEC*, No. 10-1305 (D.C. Cir. 22 July 2011).

liquidity, so it emerged along the lines of a natural monopoly. After Regulation NMS, which the NYSE had endorsed, the advantage to the NYSE diminished.[25] Because of the change in regulation, many new trading platforms developed and trading execution fragmented. Clearly, the structure of regulation plays a crucial role with respect to the viability of different order tactics and even the viability of the business models underlying different trading platforms.

The history of the money market mutual fund industry is another interesting example of how regulation can affect business models. Money market mutual funds in the United States first arose in response to Regulation Q in the early 1970s, which imposed a ceiling on the interest rates paid by banks for various types of bank deposits. When market interest rates rose above the ceiling, there was considerable migration from bank deposits toward marketed fixed-income instruments, such as Treasury bills and notes. Money market mutual funds developed in response to the binding Regulation Q rate ceilings. During the 2008 global financial crisis, the collapse of a major U.S. money market mutual fund (Reserve Fund) led to a run until the government launched a short-term insurance program to protect money market mutual fund balances. Government policy (motivated by an attempt to stabilize the financial system) helped to protect this product. However, in response to resulting pressures from banks and the new advantage that the money market fund industry obtained, the Federal Deposit Insurance Corporation then raised its insurance limit to $250,000 from $100,000. As this example illustrates, regulatory constraints have played a major role in the organization of short-term deposits in the United States. Changes in the effective regulatory structure have led to dramatic changes in the competitive landscape.

Another interesting class of issues occurs with respect to the allocation of the pricing of joint products. For example, it can be difficult to separate fully the underlying economics associated with the production, transmission, and distribution of energy products. Government regulation can affect the structure of the industry. For example, suppose there is a natural monopoly with respect to the distribution stage. How much should the provider of those services be able to obtain from the consumer or other companies providing upstream services, such as an energy product or access to a communication network? While for some products there is increased and vigorous competition, these issues are still very important with respect to the returns available from building various infrastructure components. Although the market can sort out the allocation of profits and pricing across stages when there is vigorous competition at each stage, these issues are difficult in the face of a natural monopoly. Of course, monopoly power is at the root of one of the most important traditional uses of regulation—to set pricing and returns at utility providers. In many areas, a government regulator sets public utility prices because a utility provider has a monopolistic position.

At a global level, there has been an extensive debate about *network neutrality*. Advocates of the principle of network neutrality argue that there should not be any restriction on access to networks (such as telephone and the Internet) and assert that operators would otherwise attempt to restrict competition and create market power. Conversely, some ability to discriminate on the part of service providers may be desirable for developing and supporting the underlying infrastructure. The issues involving network neutrality are not fully resolved and likely to be ones of continuing conflict. Alternative government policies may lead to very different incentives and structures in the industry.

It also is interesting to reflect on industries or sectors that receive subsidies from government policy compared with those that pay taxes. This reflection provides a means for

[25]See the discussion of the impact of Regulation NMS in Angel, Harris, and Spatt (2011).

understanding the extent to which government policy leads some sectors of the economy to be smaller and some sectors to be larger. For example, the policies of many governments have led the cigarette industry to shrink. Cigarette products often are heavily taxed, thereby reducing the size of the sector. Governments justify the heavy taxation based on externalities that are created by cigarette smoking (these include secondhand smoke and elevated health care costs for society). In contrast, many governments provide government-sponsored health care programs or heavily subsidized health care, which may increase the size of the health care sector. Because of government programs, the ultimate user pays directly only a small portion, if any, of the health care costs attributable to the user. Because the cost is not visible to or paid directly by the user, these health care policies may have the effect of encouraging the allocation of additional resources to the health care sector.

While there may be justifications for policies that heavily tax or subsidize a sector, the link between government policies and the allocation of resources to a sector can be significant and should be considered. Once again, the importance and relevance of cost-benefit analysis is apparent.

6.1. Effects of Regulations

Some regulation is very specific and focused on a particular sector, whereas other regulation is wide-ranging and may affect a number of sectors to varying degrees. Examples of regulations that focus on a particular sector are those focused on the financial sector. Financial institutions, including banks, are subject to extensive regulation from a variety of regulatory bodies. This regulation affects a variety of items, including a bank's financial and operating structures and risk management. A major regulatory challenge confronting banking supervisors is how much capital should be required for banks and other financial institutions. This issue is key for bank regulators globally, including central bankers and the Basel Committee on Banking Supervision. Increasing capital requirements may improve liquidity and stability of financial institutions and reduce reliance on indirect governmental guarantees and support.[26] Such proposals, however, have been controversial with the leadership of financial institutions, who view equity capital as substantially more costly than debt financing (this subject is discussed further later in this subsection). They argue that such regulation may reduce access to credit and hinder economic growth.

One of the lessons from the 2008 global financial crisis was that globally, many major financial institutions, including financial services firms, had excessive leverage. This leverage reflected the similarity in the exposures of many financial institutions, a factor in the spreading of losses from institution to institution in the financial system (contagion). Some of these institutions ultimately collapsed or received some sort of a bailout from a government. The similarity in exposures makes it doubtful that the financial difficulties simply reflected externalities rather than similar and apparently mistaken judgments.

An apparent conclusion of policy makers in the United States after the collapse of Lehman Brothers was that the imposition of losses on creditors would lead to a loss of confidence and systemic failure. Consequently, many participants in the financial system anticipate that bondholders of major companies are likely to be protected in the future, leading to limits on the extent that credit spreads can serve as a disciplinary mechanism and amplifying the willingness of companies to assume greater risks. The extent of this moral hazard may have been amplified

[26]Capital requirements are discussed in detail by Admati, DeMarzo, Hellwig, and Pfleiderer (2010).

because of what market participants inferred from government responses during the 2008 global financial crisis, which has manifested itself in terms of a funding advantage to banks that have greater systemic risk. To the degree that these institutions are largely debt-financed rather than equity-financed, the cost of the company's risk bearing is understated (because it is subsidized by government) and the company's shareholders may not adequately internalize the company's risk taking.

The responses of governments to, in effect, subsidize certain investors in financial institutions and the expected effects on moral hazard and on taxpayers are not unique to the United States. The Irish government guaranteed the liabilities of Irish banks at significant cost to Irish taxpayers. In 2008, Dexia, a Franco-Belgian financial services group, received a bailout from the French and Belgian governments. As reported in October 2011 in the *Wall Street Journal*, "The French and Belgian governments, both part-owners of Dexia thanks to the previous bailout, committed Tuesday to safeguarding the bank's depositors and creditors. Even if this support calms markets short-term, the maneuver poses the danger of transferring more risk from the private sector to the public one."[27] It is very challenging to assess the costs and benefits of these decisions to the various stakeholders, including taxpayers.

In countries in which interest on debt, but not dividends on equity, is tax-deductible, taxpayers effectively subsidize the use of corporate debt as a source of capital. Requirements that financial institutions meet specified capital standards are essentially requirements that they have a significant portion of equity capital so that they bear most of the marginal funding cost. It is important to recognize what high capital standards are not—they are not a requirement that companies set aside as a buffer and hold (e.g., idle) amounts of capital as a reserve. High capital standards are an attempt to ensure that investors in the financial institutions are bearing the costs of the risk that the financial institutions choose to assume.

Under the Modigliani–Miller capital structure theory, in the absence of taxes and given certain other assumptions, equity and debt are substitutes for financing the company, and capital structure is irrelevant to the value of the company.[28] To the extent that a company can reduce its overall funding cost by reducing its equity investment and increasing leverage, it reflects a government subsidy in the form of the tax advantage of debt. The addition of implicit or assumed guarantees, that the government will not allow the institution to fail, encourages increased leverage because this guarantee is in effect a government subsidy that will reduce overall funding costs. Some CEOs of financial institutions have not been sympathetic to high equity capital standards; this disagreement may reflect compensation arrangements focused on return on equity (ROE) rather than return on assets (ROA) and benefits of leverage due to the tax advantage of debt.

The bank supervision process is another aspect of regulation with important implications for the riskiness of a bank's contingent claims. Traditionally, supervision has had limited transparency; ratings and recommendations of bank supervisors have not been shared publicly. The stress tests undertaken in 2009 in the United States were cited as a major advance with respect to the bank supervision model (for example, see Bernanke 2009). In those tests, the financial institutions were subject to common shocks and stresses, which was an important innovation with respect to the supervisory model. Another key aspect of the implementation of the 2009 stress tests was the public disclosure of the funding needs of the financial institutions. Similar stress tests were and are being conducted in various parts of the world, including Europe. Issues about the disclosure of results—how much and what to

[27] *Wall Street Journal*, "Dexia Looks for Bailout, Part Deux," 5 October 2011.

[28] This is discussed in greater detail in Aggarwal, Drake, Kobor, and Noronha (2011).

disclose—and the adequacy of the tests are among the issues that regulators have had to address. Decisions about the appropriate amount of disclosure have varied among regulators.

Another cost-benefit consideration is the regulatory impact on funding costs in the capital market when a regulator essentially "infuses capital or writes checks." Although it is not common for a regulator to "write checks" (most do not have resources for such purposes), such a situation can arise with discount window borrowing from a central bank.[29] Discount window borrowers are likely to advocate that discount window transactions be kept confidential; they hope to avoid the adverse stigma associated with acknowledging their need to obtain lender of last resort financing.[30] Often, a central bank also supports this position of confidentiality; the central bank justifies this based on its objectives, such as maintaining confidence and stability in the financial system. Others may advocate for the release of this information because of the potential relevance to stakeholders, including customers and investors.

Judicial law may be required to resolve the issue of information availability in different parts of the world. In the United States, Bloomberg LP brought a Freedom of Information Act lawsuit to require disclosure of the discount window borrowers during the 2008 global financial crisis. The Federal Reserve fought the lawsuit unsuccessfully. Even with a lack of public disclosure about discount window borrowing, there is at least some empirical evidence suggesting that the marketplace penalizes financial institutions that borrow at the window (see Armantier, Ghysels, Sarkar, and Shrader 2011). It is plausible that, even in the absence of disclosure, participants in the financial markets are able to deduce that such borrowing has or is likely to occur; discount window borrowing is not the only information about an institution's financial condition available to market participants.

A final example of a regulatory issue with significant ramifications for funding costs concerns the extent to which regulators outsource the determination of regulatory treatment to rating agencies. Regulators may reference and rely on third-party ratings for regulatory purposes, such as capital requirements for financial institutions and suitability standards for particular investors or products. Ratings have an important impact on the pricing of bonds and the structuring of portfolios—for example, what bonds can be held in various types of accounts and funds (such as money market mutual funds). Those seeking ratings may engage in regulatory arbitrage (entities identify and use some aspect of regulations that allows them to exploit differences in economic substance and regulatory interpretation to their benefit) by shopping for a higher rating. The default rate of highly rated mortgage market financial instruments played an important role in the 2008 global financial crisis.

The collateralized mortgage-backed security (MBS) structure is based on the securitization model of buying a pool of assets (mortgages, in this case) and assigning the income and principal returns into individual security tranches. It was subsequently found that in many cases a pool of mortgages was transformed via securitization from individual mortgages with a lower average rating to a securitized pool of mortgages with a higher rating. The credit-rating agencies assigned the ratings to MBSs, and the ratings led to false comfort by investors about

[29]Although rare, other examples exist of regulators having resources to assist those they regulate. An example is the Troubled Asset Relief Program (TARP) that was approved in the United States in late 2008. TARP in essence provided a "checkbook" to the Secretary of the Treasury. In the United States, access to resources potentially explains the power of the Federal Reserve compared with that of the SEC or Commodity Futures Trading Commission (CFTC) during a financial crisis.

[30]A recent study documents that borrowers are willing to pay a premium to access other Federal Reserve liquidity facilities rather than the discount window. See Armantier, Ghysels, Sarkar, and Shrader (2011).

the quality of their holdings. Many observers believed that the mistakes of credit-rating agencies and the process and standards implicitly sanctioned by regulators led to the development of a relatively common view. Considerable revaluation and systemic risk ensued when the degree of risk and quality of the underlying mortgage pools did not match the higher ratings assigned by the rating agencies.

Because credit ratings were identified as a key source of systemic risk in the United States, the Dodd-Frank Act barred references to ratings in regulation. Globally, other regulators are reconsidering the roles and/or types of rating agencies; in the EU, some have advocated the creation of an EU-sponsored independent credit-rating agency. At the same time, regulators have struggled with how to replace ratings for regulatory purposes. For example, the use of credit default swap prices or yield spreads would be forward-looking, but the underlying markets may lack sufficient liquidity or could be subject to manipulation. Even the prices of the bonds could be manipulated. Although it is generally not referred to as manipulation, the large-scale purchases of European sovereign debt by the European Central Bank (ECB) can be viewed as distorting market pricing.[31] Transferring authority to organizations other than credit-rating agencies may not result in fundamental changes. It is difficult to identify an alternative approach that would eliminate the potential for systemic risk.

For some purposes, it may be helpful to have asset managers take responsibility for the classification and rating of the assets that they purchase (for example, develop suitability standards for the relevant account), but it is unclear how this approach could work with respect to capital standards. The discussion of the complexity of setting up a suitable system of capital regulation and the viability of risk-based models suggest that alternatives to the use of ratings by credit-rating agencies still may lead to systemic risk.

Despite the challenges with the credit-rating agency model, several aspects are worth emphasizing. While the discussion highlights some of the consequences in outsourcing the determination of regulatory treatment to rating agencies, to a degree there are similar challenges with many of the proposed alternatives. Reliance on intermediaries with a stake in the economic outcomes of the entity being assessed may align incentives of the assessor with those relying on the assessment. This approach, however, is not without potential conflict of interest issues. A possible benefit of the rating agency approach is that the rating agencies have the potential to provide objective guidance for financial instruments in regulatory contexts. The potential economies of scale in information production also suggest that there may be significant private incentives to outsource a portion of the information production that asset managers require. Even if regulators eliminate reference to ratings by credit-rating agencies, there may still be a viable role and business model for credit-rating agencies.

Environmental, property rights, and labor regulations are other examples of areas in which regulation plays an important role with a significant impact on society. Regulation in any of these areas may still focus on a particular industry. When analyzing an industry or business, it is important to consider the type of regulation that the industry is sensitive and susceptible to. For example, oil, gas, and mining companies, as well as certain types of manufacturers may be more sensitive to changes in the regulatory atmosphere with respect to

[31]For example, C. Spatt and P. Wallison, "A Regulatory Blueprint for Mismanaging the Sovereign Debt Crisis," Shadow Financial Regulatory Committee Statement No. 320 (2011), points to this distortion. That discussion also highlights a range of other ways in which sovereign debtors have been unhappy with various market institutions (such as credit default swaps, short selling, and credit-rating agencies) and have attempted to diminish how sovereign debt would be disciplined by markets.

environmental issues. Labor-intensive industries may be affected to a greater extent than a capital-intensive industry by regulatory changes with respect to labor conditions and rights. Pharmaceutical and technology companies may be sensitive to regulations with respect to intellectual property rights. Having a framework within which to consider regulation is very useful.

7. SUMMARY

Knowledge of regulation is important because regulation has potentially far-reaching and significant effects. These effects can range from macro-level effects on the economy to micro-level effects on individual entities and securities.

Regulation originates from a variety of sources and in a variety of areas. A framework that includes types of regulators and regulation as well as areas of regulation that may affect the entity of interest (including the economy as an entity) is useful. The framework will help in assessing possible effects of new regulation. It can also help in assessing the effects of regulation on various entities.

More than one regulator may develop regulations in response to a particular issue. Each of the relevant regulators may have a different objective and choose to address the issue using different regulatory tools.

In developing regulations, the regulator should consider costs and benefits. In the analysis, the net regulatory burden (private costs less private benefits of regulation) may also be relevant. Costs and benefits, regardless of the perspective, may be difficult to assess. A critical aspect of regulatory analysis, however, is assessing the costs and benefits of regulation.

Some key points of the chapter are summarized here:

- Legislative bodies, regulatory bodies, and courts typically enact regulation.
- Regulatory bodies include government agencies and independent regulators granted authority by a government or governmental agency. Some independent regulators may be self-regulating organizations.
- Typically, legislative bodies enact broad laws or statutes; regulatory bodies issue administrative regulations, often implementing statutes; and courts interpret statutes and administrative regulations, and these interpretations may result in judicial law.
- Regulators have responsibility for both substantive and procedural laws. The former focuses on rights and responsibilities of entities and relationships among entities. The latter focuses on the protection and enforcement of the former.
- The existence of informational frictions and externalities creates a need for regulation. Regulation is expected to have societal benefits and should be assessed using cost-benefit analysis.
- Regulation that arises to enhance the interests of regulated entities reflects regulatory capture.
- Regulatory competition is competition among different regulatory bodies to use regulation in order to attract certain entities.
- Regulatory arbitrage is the use of regulation by an entity to exploit differences in economic substance and regulatory interpretation or in regulatory regimes to the entity's benefit.
- Interdependence in the actions and potentially conflicting objectives of regulators is an important consideration for regulators, those regulated, and those assessing the effects of regulation.

- There are many regulatory tools available to regulators, including price mechanisms (such as taxes and subsidies), regulatory mandates and restrictions on behaviors, provision of public goods, and public financing of private projects.
- The choice of regulatory tool should be consistent with maintaining a stable regulatory environment. Stable does not mean unchanging, but rather refers to desirable attributes of regulation, including predictability, effectiveness in achieving objectives, time consistency, and enforceability.
- The breadth of regulation of commerce necessitates the use of a framework that identifies potential areas of regulation. This framework can be referenced to identify specific areas of regulation, existing and anticipated, that may affect the entity of interest.
- The regulation of securities markets and financial institutions is extensive and complex because of the consequences of failures in the financial system. These consequences include financial losses, loss of confidence, and disruption of commerce.
- The focus of regulators in financial markets includes prudential supervision, financial stability, market integrity, and economic growth, among others.
- Regulators—in assessing regulation and regulatory outcomes—should conduct ongoing cost-benefit analyses, develop techniques to enhance the measurement of these analyses, and use economic principles for guidance.
- Net regulatory burden to the entity of interest is an important consideration for an analyst.

PRACTICE PROBLEMS

The following information relates to Questions 1 through 6.[32]

Tiu Asset Management (TAM) recently hired Jonna Yun. Yun is a member of TAM's Global Equity portfolio team and is assigned the task of analyzing the effects of regulation on the U.S. financial services sector. In her first report to the team, Yun makes the following statements:

Statement 1: "The Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act), enacted on 21 July 2010 by the U.S. Congress, will have a significant effect on U.S. banks and other financial services firms."

Statement 2: "The U.S. Securities and Exchange Commission (SEC) allocates certain regulatory responsibilities to the Financial Industry Regulatory Authority (FINRA), with the goal of ensuring that the securities industry operates fairly and honestly."

Statement 3: "The Dodd-Frank Act called for derivatives reforms, including shifting from bilateral to centralized derivatives settlement, by July 2011. The G-20 called for action by its members on derivatives reform by year-end 2012. The accelerated time line of the Dodd-Frank Act concerned some U.S. firms."

Statement 4: "Regulators use various tools to intervene in the financial services sector."

Statement 5: "Regulations may bring benefits to the U.S. economy, but they may also have unanticipated costly effects."

Statement 6: "Regulation Q imposed a ceiling on interest rates paid by banks for certain bank deposits."

[32]These practice problems were developed by E. Shepard Farrar, CFA (Washington, D.C., USA).

1. The *most* appropriate classification of the Dodd-Frank Act, referred to in Statement 1, is a(n):
   A. statute.
   B. judicial law.
   C. administrative law.

2. The Financial Industry Regulatory Authority, referred to in Statement 2, is *best* classified as a:
   A. legislative body.
   B. government agency.
   C. self-regulatory organization.

3. What is the *most likely* basis for the concerns noted in Statement 3?
   A. Externalities
   B. Regulatory arbitrage
   C. Informational friction

4. The tools *least likely* to be used by regulators to intervene in financial markets are:
   A. blackout periods.
   B. capital requirements.
   C. insider trading restrictions.

5. Which of the following is *most likely* an unanticipated effect of regulation?
   A. Hiring compliance lawyers
   B. Setting legal standards for contracts
   C. Establishing employers' rights and responsibilities

6. After Regulation Q was imposed, the demand for money market funds *most likely*:
   A. increased.
   B. decreased.
   C. remained unchanged.

GLOSSARY

Abnormal profit *See* Economic profit.

Absolute advantage The situation in which a country is able to produce a good at a lower cost or use fewer resources in its production than a trading partner.

Absolute convergence The idea that developing countries, regardless of their particular characteristics, will eventually catch up with the developed countries and match them in per capita output.

Absolute version of PPP The extension of the law of one price to the broad range of goods and services that are consumed in different countries.

Accounting (or explicit) costs Payments to nonowner parties for services or resources they supply to the firm.

Accounting loss When accounting profit is negative.

Accounting profit Income before taxes or pretax income; income as reported on the income statement, in accordance with prevailing accounting standards, before the provisions for income tax expense.

Action lag A delay from policy decisions to implementation.

Activity ratio Also called the *participation ratio*, this is the ratio of the labor force to total population of working age.

Administrative regulations or **administrative law** Rules issued by government agencies or other regulators.

Aggregate demand The quantity of goods and services that households, businesses, government, and foreign customers want to buy at any given level of prices.

Aggregate demand curve Inverse relationship between the price level and real output.

Aggregate income The value of all the payments earned by the suppliers of factors used in the production of goods and services.

Aggregate output The value of all the goods and services produced in a specified period of time.

Aggregate supply The quantity of goods and services producers are willing to supply at any given price level.

Aggregate supply curve The level of domestic output that companies will produce at each price level.

Autarkic price The price of a good or service in a country that does not trade with other countries.

Autarky The state in which a country does not trade with other countries.

Automatic stabilizer A countercyclical factor that automatically comes into play as an economy slows and unemployment rises.

Average fixed cost Total fixed cost divided by quantity.

Average product Measures the productivity of inputs on average and is calculated by dividing total product by the total number of units for a given input that is used to generate that output.

Average revenue Quantity sold divided into total revenue.

Average total cost Total costs divided by quantity.

Average variable cost Total variable cost divided by quantity.

Balanced With respect to a government budget, one in which spending and revenues (taxes) are equal.

Balance of payments A double-entry bookkeeping system that summarizes a country's economic transactions with the rest of the world for a particular period of time, typically a calendar quarter or year.

Balance of trade deficit When the domestic economy is spending more on foreign goods and services than foreign economies are spending on domestic goods and services.

Barter economy An economy where economic agents as households, corporations, and governments pay for goods and services with another good or service.

Base rate The reference rate on which a bank bases lending rates to all other customers.

Bond market vigilantes Bond market participants who might reduce their demand for long-term bonds, thus pushing up their yields.

Boom An expansionary phase characterized by economic growth testing the limits of the economy.

Breakeven point The number of units produced and sold at which the company's net income is zero (revenues equal total costs); in the case of perfect competition, the quantity where price, average revenue, and marginal revenue equal average total cost.

Broad money Encompasses narrow money plus the entire range of liquid assets that can be used to make purchases.

Budget constraint A constraint on spending or investment imposed by wealth or income.

Budget surplus/deficit The difference between government revenue and expenditure for a stated fixed period of time.

Capital account A component of the balance of payments that measures transfers of capital.

Capital consumption allowance A measure of the wear and tear (depreciation) of the capital stock that occurs in the production of goods and services.

Capital deepening An increase in the capital-to-labor ratio.

Capital-deepening investment Increases the stock of capital relative to labor.

Capital expenditure Expenditure on physical capital (fixed assets).

Capital restrictions Controls placed on foreigners' ability to own domestic assets and/or domestic residents' ability to own foreign assets.

Capital stock The accumulated amount of buildings, machinery, and equipment used to produce goods and services.

Cartel Participants in collusive agreements that are made openly and formally.

Central bank The dominant bank in a country, usually with official or semiofficial governmental status.

Closed economy The economy within a country that does not trade with other countries (*see* Autarky).

Club convergence The idea that only rich and middle-income countries sharing a set of favorable attributes (i.e., they are members of the "club") will converge to the income level of the richest countries.

Cobb–Douglas production function A function of the form $Y = K^{\alpha}L^{1-\alpha}$ relating output (Y) to labor (L) and capital (K) inputs.

Coincident economic indicators Turning points that are usually close to those of the overall economy; they are believed to have value for identifying the economy's present state.

Common market A regional trading bloc that incorporates all aspects of a customs union and extends it by allowing free movement of factors of production among members.

Comparative advantage The situation in which a country faces an opportunity cost of producing a good that is less than that of its trading partner.

Complements Said of goods that tend to be used together; technically, two goods whose cross-price elasticity of demand is negative.

Complete preferences The assumption that a consumer is able to make a comparison between any two possible bundles of goods.

Conditional convergence The idea that convergence of per capita income is conditional on the countries having the same savings rate, population growth rate, and production function.

Conspicuous consumption Consumption of high-status goods, such as a luxury automobile or a very expensive piece of jewelry.

Constant-cost industry When firms in the industry experience no change in resource costs and output prices over the long run.

Constant returns to scale The characteristic of constant per-unit costs in the presence of increased production; the condition that if all inputs into the production process are increased by a given percentage, then output rises by that same percentage.

Consumer choice theory The theory relating consumer demand curves to consumer preferences.

Consumer surplus The difference between the value that a consumer places on units purchased and the (smaller) amount of money that was required to pay for them.

Consumption The demand for goods and services.

Consumption basket or **consumption bundle** A specific combination of the goods and services that a consumer wants to consume.

Contraction The period of a business cycle after the peak and before the trough; often called a *recession* or, if exceptionally severe, called a *depression*.

Contractionary Tending to cause the real economy to contract.

Contractionary fiscal policy A fiscal policy that has the objective to make the real economy contract.

Convergence The tendency for differences in output per capita across countries to diminish over time; in technical analysis, a term that describes the case when an indicator moves in the same manner as the security being analyzed.

Core inflation The inflation rate calculated based on a price index of goods and services except food and energy.

Cost-push The type of inflation in which rising costs, usually wages, compel businesses to raise prices generally.

Cournot assumption Assumption in which each firm determines its profit-maximizing production level, assuming that the other firms' outputs will not change.

Covered interest rate parity Relationship among the spot exchange rate, forward exchange rate, and the interest rates in two currencies that ensures that the return on a hedged (i.e., covered) foreign risk-free investment is the same as the return on a domestic risk-free investment.

Cross-price elasticity of demand The percent change in quantity demanded for a given small change in the price of another good; the responsiveness of the demand for product A that is associated with the change in price of product B.

Crowding out The thesis that government borrowing may divert private-sector investment from taking place.

Current account A component of the balance of payments that measures the flow of goods and services.

Current government spending Government expenditures on goods and services that are provided on a regular, recurring basis, including health, education, and defense.

Customs union A regional trading bloc that allows free movement of goods and services among members while also creating a common trade policy against nonmembers.

Cyclical companies Companies with sales and profits that regularly expand and contract with the business cycle or state of the economy.

Decreasing-cost industry An industry in which per-unit costs and output prices are lower when industry output is increased in the long run.

Decreasing returns to scale Increase in cost per unit resulting from increased production.

Defensive companies Companies with sales and profits that have little sensitivity to the business cycle or state of the economy.

Deflation Negative inflation.

Demand-pull Type of inflation in which increasing demand raises prices generally, which then are reflected in a business's costs as workers demand wage hikes to catch up with the rising cost of living.

Demand and supply analysis The study of how buyers and sellers interact to determine transaction prices and quantities.

Demand shock A typically unexpected disturbance to demand, such as an unanticipated interruption in trade or transportation.

Depression *See* Contraction.

Diffusion index Reflects the proportion of an index's components that are moving in a pattern consistent with the overall index.

Diminishing marginal productivity Describes a state in which each additional unit of an input, keeping the other inputs unchanged, increases output by a smaller increment.

Direct taxes Taxes levied directly on income, wealth, and corporate profits.

Discount rate With reference to U.S. banking, the rate for member banks borrowing directly from the U.S. Federal Reserve System.

Discouraged worker A person who has stopped looking for a job or has given up seeking employment.

Diseconomies of scale Increase in cost per unit resulting from increased production.

Disinflation A decline in the inflation rate, such as from around 15 to 20 percent to 5 or 6 percent.

Domestic content provisions Stipulations that some percentage of the value added or components used in production should be of domestic origin.

Double coincidence of wants A prerequisite to barter trades, in particular that both economic agents in the transaction want what the other is selling.

Dutch disease A situation in which currency appreciation driven by strong export demand for resources makes other segments of the economy (particularly manufacturing) globally uncompetitive.

Economic cost All the remuneration needed to keep a productive resource in its current employment or to acquire the resource for productive use; the sum of total accounting costs and implicit opportunity costs.

Economic indicator A variable that provides information on the state of the overall economy.

Economic loss The amount by which accounting profit is less than normal profit.

Economic profit Also called *abnormal profit* or *supernormal profit*, it is equal to accounting profit less the implicit opportunity costs not included in total accounting costs; the difference between total revenue (TR) and total cost (TC).

Economic rent The surplus value that results when a particular resource or good is fixed in supply and the market price is higher than what is required to bring the resource or good onto the market and sustain its use.

Economic union A regional trading bloc that incorporates all aspects of a common market and additionally requires common economic institutions and coordination of economic policies among members.

Economics The study of production, distribution, and consumption. Economics is divided into two broad areas of study: macroeconomics and microeconomics.

Economic stabilization Reduction of the magnitude of economic fluctuations.

Economies of scale Reduction in cost per unit resulting from increased production.

Elasticity of supply A measure of the sensitivity of quantity supplied to a change in price.

Employed The number of people with a job.

***Ex ante* version of PPP** Hypothesis that expected changes in the spot exchange rate are equal to expected differences in national inflation rates. An extension of relative purchasing power parity to expected future changes in the exchange rate.

Expansion The period of a business cycle after its lowest point and before its highest point.

Expansionary Tending to cause the real economy to grow.

Expansionary fiscal policy Fiscal policy aimed at achieving real economic growth.

Expected inflation The level of inflation that economic agents expect in the future.

Exports Goods and services that a domestic economy sells to other countries.

Export subsidy A payment by a government to a firm that exports a good that is being subsidized.

Externalities Spillover effects of production and consumption activities onto others who did not consent to participate in the activity.

External sustainability approach An approach to assessing the equilibrium exchange rate that focuses on exchange rate adjustments required to ensure that a country's net foreign-asset/GDP ratio or net foreign-liability/GDP ratio stabilizes at a sustainable level.

Federal funds rate or **fed funds rate** The U.S. interbank lending rate on overnight borrowings of reserves.

Fiat money Money that is not convertible into any other commodity.

Financial account A component of the balance of payments that records investment flows.

Financial contagion A situation where financial shocks spread from their place of origin to other locales; in essence, a faltering economy infects other, healthier economies.

First-degree price discrimination Where a monopolist is able to charge each customer the highest price the customer is willing to pay.

Fiscal multiplier The ratio of a change in national income to a change in government spending.

Fiscal policy The use of taxes and government spending to affect the level of aggregate expenditures.

Fisher effect The thesis that the real rate of interest in an economy is stable over time so that changes in nominal interest rates are the result of changes in expected inflation.

Fisher index The geometric mean of the Laspeyres index.

Foreign currency reserves Holdings by the central bank of nondomestic currency deposits and nondomestic bonds.

Foreign direct investment The direct investment by a firm domiciled in one country in productive assets in a foreign country.

Foreign portfolio investment Investment by individuals, firms, and institutional investors in foreign financial instruments.

Fractional reserve banking Banking in which reserves constitute a fraction of deposits.

Free trade When there are no government restrictions on a country's ability to trade.

Free trade areas A regional trading bloc within which all barriers to the flow of goods and services among members have been eliminated.

FX carry trade An investment strategy that involves taking on long positions in high-yield currencies and short positions in low-yield currencies.

FX swap The combination of a spot and a forward FX transaction.

Game theory The set of tools decision makers use to incorporate responses by rival decision makers into their strategies.

GDP deflator A gauge of prices and inflation that measures the aggregate changes in prices across the overall economy.

Giffen good A good that is consumed more as the price of the good rises.

Gilts Bonds issued by the UK government.

Gold standard With respect to a currency, a given amount of a currency that is on the gold standard can be converted into a prespecified amount of gold.

Gross domestic product The market value of all final goods and services produced within the economy in a given period of time (output definition) or, equivalently, the aggregate income earned by all households, all companies, and the government within the economy in a given period of time (income definition).

Growth accounting equation The production function written in the form of growth rates. For the basic Cobb–Douglas production function, it states that the growth rate of output equals the rate of technological change plus α times the growth rate of capital plus $(1 - \alpha)$ times the growth rate of labor.

Headline inflation The inflation rate calculated based on the price index that includes all goods and services in an economy.

Horizontal demand schedule Implies that at a given price the response in the quantity demanded is infinite.

Household A person or a group of people living in the same residence, taken as a basic unit in economic analysis.

Human capital The accumulated knowledge and skill that workers acquire from education, training, or life experience.

Hyperinflation An extremely fast increase in aggregate price level, which corresponds to an extremely high inflation rate—for example, 500 to 1,000 percent per year.

Impact lag The lag associated with the result of actions affecting the economy with delay.

Imperfect competition A market structure in which an individual firm has enough share of the market (or can control a certain segment of the market) such that it is able to exert some influence over price.

Implicit price deflator for GDP A gauge of prices and inflation that measures the aggregate changes in prices across the overall economy.

Imports Goods and services that a domestic economy purchases from other countries.

Income constraint The constraint on consumers to spend, in total, no more than their income.

Income elasticity of demand A measure of the responsiveness of demand to changes in income, defined as the percentage change in quantity demanded divided by the percentage change in income.

Increasing-cost industry An industry in which per-unit costs and output prices are higher when industry output is increased in the long run.

Increasing marginal returns Where the marginal product of a resource increases as additional units of that input are employed.

Increasing returns to scale Reduction in cost per unit resulting from increased production.

Independent regulators Regulators recognized and granted authority by a government body or agency. They are not government agencies per se and typically do not rely on government funding.

Index of leading economic indicators A composite of economic variables used by analysts to predict future economic conditions.

Indifference curve A curve representing all the combinations of two goods or attributes such that the consumer is entirely indifferent among them.

Indifference curve map A group or family of indifference curves, representing a consumer's entire utility function.

Indirect taxes Taxes such as taxes on spending, as opposed to direct taxes.

Inelastic Insensitive to price changes.

Inelastic supply Said of supply that is insensitive to the prices of goods sold.

Inflation The percentage increase in the general price level from one period to the next; a sustained rise in the overall level of prices in an economy.

Inflation rate The percentage change in a price index—that is, the speed of overall price level movements.

Inflation reports A type of economic publication put out by many central banks.

Inflation uncertainty The degree to which economic agents view future rates of inflation as difficult to forecast.

Informational frictions Forces that restrict availability, quality, and/or flow of information and its use.

Interest Payment for lending funds.

International Fisher effect Proposition that nominal interest rate differentials across currencies are determined by expected inflation differentials.

Inventory investment Net change in business inventory.

Judicial law Interpretations of courts.

Keynesians Economists who believe that fiscal policy can have powerful effects on aggregate demand, output, and employment when there is substantial spare capacity in an economy.

Labor force The portion of the working age population (ages 16 to 64) who either is employed or is available for work but not working (unemployed).

Labor force participation rate The percentage of the working age population that is in the labor force.

Labor productivity The quantity of goods and services (real GDP) that a worker can produce in one hour of work. More generally, output per unit of labor input.

Labor productivity growth accounting equation States that potential GDP growth equals the growth rate of the labor input plus the growth rate of labor productivity.

Lagging economic indicators Turning points that take place later than those of the overall economy; they are believed to have value in identifying the economy's past condition.

Laspeyres index A price index created by holding the composition of the consumption basket constant.

Law of diminishing returns As additional resources are added to a production process, the marginal product of each successive additional resource is lower. Eventually, an additional resource reduces the marginal product (negative marginal product).

Law of one price Hypothesis that (1) identical goods should trade at the same price across countries when valued in terms of a common currency, or (2) two equivalent financial instruments or combinations of financial instruments can sell for only one price. The latter form is equivalent to the principle that no arbitrage opportunities are possible.

Leading economic indicators Turning points that usually precede those of the overall economy; they are believed to have value for predicting the economy's future state, usually near-term.

Legal tender Something that must be accepted when offered in exchange for goods and services.

Lender of last resort An entity willing to lend money when no other entity is ready to do so.

Liquidity trap A condition in which the demand for money becomes infinitely elastic (horizontal demand curve) so that injections of money into the economy will not lower interest rates or affect real activity.

Long-run average total cost curve The curve describing average total costs when no costs are considered fixed.

Long-run industry supply curve A curve describing the relationship between quantity supplied and output prices when no costs are considered fixed.

Macroeconomic balance approach An approach to assessing the equilibrium exchange rate that focuses on exchange rate adjustments needed to close the gap between the medium-term expectation for a country's current account balance and that country's normal (or sustainable) current account balance.

Macroeconomics The area of study in economics that deals with aggregate economic quantities, such as national output and national income.

Marginal cost The cost of producing an additional unit of a good.

Marginal product Measures the productivity of each unit of input and is calculated by taking the difference in total product from adding another unit of input (assuming other resource quantities are held constant).

Marginal propensity to consume The proportion of an additional unit of disposable income that is consumed or spent; the change in consumption for a small change in income.

Marginal propensity to save The proportion of an additional unit of disposable income that is saved (not spent).

Marginal rate of substitution The rate at which one is willing to give up one good to obtain more of another.

Marginal revenue The change in total revenue divided by the change in quantity sold; simply, the additional revenue from selling one more unit.

Marginal revenue product The amount of additional revenue received from employing an additional unit of an input.

Marginal value curve A curve describing the highest price consumers are willing to pay for each additional unit of a good.

Market structure The competitive environment (perfect competition, monopolistic competition, oligopoly, and monopoly).

Measure of value A standard for measuring value; a function of money.

Medium of exchange Any asset that can be used to purchase goods and services or to repay debts; a function of money.

Menu costs A cost of inflation in which businesses constantly have to incur the costs of changing the advertised prices of their goods and services.

Microeconomics The area of study of economics that deals with markets and decision making of individual economic units, including consumers and businesses.

Minimum efficient scale The smallest output that a firm can produce such that its long-run average cost is minimized.

Minsky moment Named for Hyman Minksy, this is a point in a business cycle when, after individuals become overextended in borrowing to finance speculative investments, they start realizing that something is likely to go wrong and a panic ensues, leading to asset sell-offs.

Monetarists Economists who believe that the rate of growth of the money supply is the primary determinant of the rate of inflation.

Monetary policy Actions taken by a nation's central bank to affect aggregate output and prices through changes in bank reserves, reserve requirements, or its target interest rate.

Monetary transmission mechanism The process whereby a central bank's interest rate gets transmitted through the economy and ultimately affects the rate of increase of prices.

Monetary union When members of an economic union adopt a common currency.

Money A generally accepted medium of exchange and unit of account.

Money creation The process by which changes in bank reserves translate into changes in the money supply.

Money multiplier Describes how a change in reserves is expected to affect the money supply; in its simplest form, 1 divided by the reserve requirement.

Money neutrality The thesis that an increase in the money supply leads in the long run to an increase in the price level, while leaving real variables like output and employment unaffected.

Monopolist Said of an entity that is the only seller in its market.

Monopolistic competition Highly competitive form of imperfect competition; the competitive characteristic is a notably large number of firms, while the monopoly aspect is the result of product differentiation.

Monopoly In pure monopoly markets, there are no substitutes for the given product or service. There is a single seller, which exercises considerable power over pricing and output decisions.

Multinational corporation A firm that operates in more than one country or has subsidiary firms in more than one country.

Narrow money The notes and coins in circulation in an economy, plus other very highly liquid deposits.

Nash equilibrium When two or more participants in a noncooperative game have no incentive to deviate from their respective equilibrium strategies, given their opponent's strategies.

National income The income received by all factors of production used in the generation of final output. National income equals gross domestic product (GDP)—or, in some countries, gross national product (GNP)—minus the capital consumption allowance and a statistical discrepancy.

Natural rate of unemployment Effective unemployment rate below which pressure emerges in labor markets.

Neo-Keynesians A group of dynamic general equilibrium models that assume slow-to-adjust prices and wages.

Net exports The difference between the value of a country's exports and the value of its imports.

Net regulatory burden The private costs of regulation less the private benefits of regulation.

Net tax rate The tax rate net of transfer payments.

Network externalities The impact that users of a good, a service, or a technology have on other users of that product; it can be positive (e.g., a critical mass of users makes a product more useful) or negative (e.g., congestion makes the product less useful).

Neutral rate of interest The rate of interest that neither spurs on nor slows the underlying economy.

New classical macroeconomics An approach to macroeconomics that seeks the macroeconomic conclusions of individuals maximizing utility on the basis of rational expectations and companies maximizing profits.

New Keynesians *See* Neo-Keynesians.

Nominal GDP The value of goods and services measured at current prices.

Nonaccelerating inflation rate of unemployment *See* Natural rate of unemployment.

Nonconvergence trap A situation in which a country remains relative poor, or even falls further behind, because it fails to implement necessary institutional reforms or adopt leading technologies.

Nonrenewable resources Finite resources that are depleted once they are consumed, such as oil and coal.

Nonsatiation The assumption that consumers could never have so much of a preferred good that they would refuse any more, even if it were free; sometimes referred to as the "more is better" assumption.

Normal profit The level of accounting profit needed to just cover the implicit opportunity costs ignored in accounting costs.

Official interest rate Also called the *official policy rate* or *policy rate*; it is an interest rate that a central bank sets and announces publicly; normally the rate at which it is willing to lend money to the commercial banks.

Official policy rate *See* Official interest rate.

Oligopoly Market structure with a relatively small number of firms supplying the market.

Open economy The economy within a country that does trade with other countries.

Open market operations Activities that involve the purchase and sale of government bonds from and to commercial banks and/or designated market makers.

Operational independence A bank's ability to execute monetary policy and set interest rates in the way it thinks would best meet the inflation target.

Opportunity cost The value that investors forgo by choosing a particular course of action; the value of something in its best alternative use.

Paasche index An index formula using the current composition of a basket of products.

Payments system The system for the transfer of money.

Peak The highest point of a business cycle.

Per capita real GDP Real gross domestic product (GDP) divided by the size of the population, often used as a measure of the average standard of living in a country.

Perfect competition Also called *price taker*; it is a market structure in which the individual firm has virtually no impact on market price, because it is assumed to be a very small seller among a very large number of firms selling essentially identical products.

Personal consumption expenditures All domestic personal consumption; the basis for a price index for such consumption, called the PCE price index.

Personal disposable income Equal to personal income less personal taxes.

Personal income A broad measure of household income that includes all income received by households, whether earned or unearned; measures the ability of consumers to make purchases.

Planning horizon A time period in which all factors of production are variable, including technology, physical capital, and plant size.

Policy rate *See* Official interest rate.

Portfolio balance approach A theory of exchange rate determination that emphasizes the portfolio investment decisions of global investors and the requirement that global investors willingly hold all outstanding securities denominated in each currency at prevailing prices and exchange rates.

Portfolio demand for money *See* Speculative demand for money.

Potential GDP The level of real GDP that can be produced at full employment; measures the productive capacity of the economy—the maximum amount of output an economy can sustainably produce without inducing an increase in the inflation rate. It is the output level that corresponds to full employment with consistent wage and price expectations.

Precautionary money balances Money held to provide a buffer against unforeseen events that might require money.

Price The market price as established by the interactions of the market demand and supply factors.

Price elasticity of demand Measures the percentage change in the quantity demanded, given a percentage change in the price of a given product.

Price index Represents the average prices of a basket of goods and services.

Price stability In economics, refers to an inflation rate that is low on average and not subject to wide fluctuation.

Price takers Producers that must accept whatever price the market dictates.

Procedural law The body of law that focuses on the protection and enforcement of the substantive laws.

Producer price index Reflects the price changes experienced by domestic producers in a country.

Production function Provides the quantitative link between the level of output that the economy can produce and the inputs used in the production process.

Production opportunity frontier Curve describing the maximum number of units of one good a company can produce, for any given number of the other good(s) that it chooses to manufacture.

Productivity The amount of output produced by workers in a given period of time—for example, output per hour worked; measures the efficiency of labor.

Profit The return that owners of a company receive for the use of their capital and the assumption of financial risk when making their investments.

Promissory note A written promise to pay a certain amount of money on demand.

Prudential supervision Regulation and monitoring of the safety and soundness of financial institutions to promote financial stability, reduce systemwide risks, and protect customers of financial institutions.

Purchasing power parity (PPP) The idea that exchange rates move to equalize the purchasing power of different currencies.

Quantitative easing An expansionary monetary policy based on aggressive open market purchase operations.

Quantity or **quantity demanded** The amount of a product that consumers are willing and able to buy at each price level.

Quantity equation of exchange An expression that over a given period, the amount of money used to purchase all goods and services in an economy, $M \times V$, is equal to monetary value of this output, $P \times Y$.

Quantity theory of money Asserts that total spending (in money terms) is proportional to the quantity of money.

Quasi-fixed cost A cost that stays the same over a range of production but can change to another constant level when production moves outside of that range.

Quota A restriction on the quantity of a good that can be imported into a country, generally for a specified period of time.

Quota rents Excess profits earned by foreign producers that raise prices after a quota is imposed and earn greater profits than they would have without the quota.

Real exchange rate The relative purchasing power of two currencies, defined in terms of the actual goods and services that each can buy at prevailing national price levels and nominal exchange rates. Measured as the ratio of national price levels expressed in a common currency.

Real GDP The value of goods and services produced, measured at base year prices.

Real income Income adjusted for the effect of inflation on the purchasing power of money.

Real interest rate Nominal interest rate minus the expected rate of inflation.

Real interest rate parity The proposition that real interest rates will converge to the same level across different markets.

Recession A period during which real GDP decreases (i.e., negative growth) for at least two successive quarters, or a period of significant decline in total output, income, employment, and sales usually lasting from six months to a year.

Recognition lag The lag in government response to an economic problem resulting from the delay in confirming a change in the state of the economy.

Refinancing rate A type of central bank policy rate.

Regulatory arbitrage Entities may identify and use some aspect of regulations that allows them to exploit differences in economic substance and regulatory interpretation or in foreign and domestic regulatory regimes to their own advantage.

Regulatory burden The costs of regulation for the regulated entity.

Regulatory capture Theory that regulation often arises to enhance the interests of the regulated.

Regulatory competition Regulators may compete to provide a regulatory environment designed to attract certain entities.

Relative price The price of a specific good or service in comparison with prices of other goods and services.

Relative version of PPP Hypothesis that changes in (nominal) exchange rates over time are equal to national inflation rate differentials.

Renewable resources Resources that can be replenished, such as forests.

Rent Payment for the use of property.

Rental price of capital The cost per unit of time to rent a unit of capital.

Repo rates Short-term collateralized lending rates.

Repurchase (repo) agreement The sale of securities together with an agreement for the seller to buy back the securities at a later date at a higher price. Typically it is a short-term agreement; if long-term, it is called a term repo.

Reserve requirement The requirement for banks to hold reserves in proportion to the size of deposits.

Ricardian equivalence An economic theory that implies that it makes no difference whether a government finances a deficit by increasing taxes or issuing debt.

Risk premium An extra return expected by investors for bearing some specified risk.

Risk reversal An option position that consists of the purchase of an out-of-the-money call and the simultaneous sale of an out-of-the-money put with the same delta, on the same underlying currency or security, and with the same expiration date.

Say's law Named for French economist J. B. Say, this law states that all that is produced will be sold because supply creates its own demand.

Second-degree price discrimination When the monopolist charges different per-unit prices using the quantity purchased as an indicator of how highly the customer values the product.

Self-regulating organizations Private, nongovernmental organizations that both represent and regulate their members. Some self-regulating organizations are also independent regulators.

Shareholder wealth maximization The process of maximizing the market value of shareholders' equity.

Short-run average total cost curve The curve describing average total costs when some costs are considered fixed.

Short-run supply curve The section of the marginal cost curve that lies above the minimum point on the average variable cost curve.

Shutdown point The point at which average revenue is less than average variable cost.

Speculative demand for money Also called the *portfolio demand for money*; the demand to hold speculative money balances based on the potential opportunities or risks that are inherent in other financial instruments.

Speculative money balances Monies held in anticipation that other assets will decline in value.

Stackelberg model A prominent model of strategic decision making in which firms are assumed to make their decisions sequentially.

Stagflation When a high inflation rate is combined with a high level of unemployment and a slowdown of the economy.

Statutes Laws enacted by legislative bodies.

Steady state rate of growth The constant growth rate of output (or output per capita) that can or will be sustained indefinitely once it is reached. Key ratios, such as the capital-to-output ratio, are constant on the steady state growth path.

Sterilized intervention A policy measure in which a monetary authority buys or sells its own currency to mitigate undesired exchange rate movements and simultaneously offsets the impact on the money supply with transactions in other financial instruments (usually money market instruments).

Store of value The quality of tending to preserve value.

Store of wealth Goods that depend on both the fact that they do not perish physically over time and on the belief that others will always value the good.

Structural (or cyclically adjusted) budget deficit The deficit that would exist if the economy was at full employment (or full potential output).

Substantive law The body of law that focuses on the rights and responsibilities of entities and relationships among entities.

Substitutes Said of two goods or services such that if the price of one increases the demand for the other tends to increase, holding all other things equal (e.g., butter and margarine).

Supernormal profit *See* Economic profit.

Supply shock A typically unexpected disturbance to supply.

Sustainable rate of economic growth The rate of increase in the economy's productive capacity or potential GDP.

Systemic risk The risk of failure of the financial system.

Target independent A central bank's ability to determine the definition of inflation that it targets, the rate of inflation that it targets, and the horizon over which the target is to be achieved.

Tariffs Taxes that a government levies on imported goods.

Technology The process a company uses to transform inputs into outputs.

Terms of trade The ratio of the price of exports to the price of imports as represented by export and import price indexes.

Theory of the consumer The branch of microeconomics that deals with consumption—the demand for goods and services—by utility-maximizing individuals.

Theory of the firm The branch of microeconomics that deals with the supply of goods and services by profit-maximizing firms.

Third-degree price discrimination When the monopolist segregates customers into groups based on demographic or other characteristics and offers different pricing to each group.

Total costs The summation of all costs, where costs are classified according to fixed or variable.

Total factor productivity (TFP) A multiplicative scale factor that reflects the general level of productivity or technology in the economy. Changes in total factor productivity generate proportional changes in output for any input combination. It is a scale factor that reflects the portion of growth that is not accounted for by explicit factor inputs (e.g., capital and labor).

Total fixed cost The summation of all expenses that do not change when production varies.

Total product The aggregate sum of production for the firm during a time period.

Total revenue Price times the quantity of units sold.

Total variable cost The summation of all variable expenses.

Trade deficit When the value of a country's exports is less than the value of its imports.

Trade diversion When lower-cost imports from nonmember countries are replaced with higher-cost imports from members.

Trade protection Government policies that impose restrictions on trade, such as tariffs and quotas.

Trade surplus When the value of a country's exports is greater than the value of its imports.

Transactions money balances Money balances that are held to finance transactions.

Transfer payments Welfare payments made through the social security system that exist to provide a basic minimum level of income for low-income households.

Transitive preferences The assumption that when comparing any three distinct bundles, A, B, and C, if A is preferred to B and simultaneously B is preferred to C, then it must be true that A is preferred to C.

Treasury Inflation-Protected Security A bond issued by the U.S. Treasury Department that is designed to protect the investor from inflation by adjusting the principal of the bond for changes in inflation.

Triangular arbitrage An arbitrage transaction involving three currencies that attempts to exploit inconsistencies among pairwise exchange rates.

Trough The lowest point of a business cycle.

Two-week repo rate The interest rate on a two-week repurchase agreement; may be used as a policy rate by a central bank.

Uncovered interest rate parity The proposition that the expected return on an uncovered (i.e., unhedged) foreign currency (risk-free) investment should equal the return on a comparable domestic currency investment.

Underemployed A person who has a job but also has the qualifications to work a significantly higher-paying job.

Unemployed A person who is actively seeking employment but is currently without a job.

Unemployment rate The ratio of unemployed to the labor force.

Unexpected (unanticipated) inflation The component of inflation that is a surprise.

Unit labor cost The average labor cost to produce one unit of output.

Unsterilized intervention A policy measure in which a monetary authority buys or sells its own currency to mitigate undesired exchange rate movements and does not offset the impact on the money supply with transactions in other financial instruments.

Util A unit of utility.

Utility function A mathematical representation of the satisfaction derived from a consumption basket.

Veblen good A good that increases in desirability along with price.

Vertical demand schedule Implies that some fixed quantity is demanded, regardless of price.

Voluntarily unemployed A person voluntarily outside the labor force, such as a jobless worker refusing an available vacancy.

Voluntary export restraint The agreement of a country to limit its exports of a good to a trading partner to a specific number of units.

Wealth effect An increase (or decrease) in household wealth increases (or decreases) consumer spending out of a given level of current income.

Wholesale price index Reflects the price changes experienced by domestic producers in a country.

REFERENCES

Admati, Anat, Peter DeMarzo, Martin Hellwig, and Paul Pfleiderer. 2010. "Fallacies, Irrelevant Facts and Myths in the Discussion of Capital Regulation: Why Bank Equity Is *Not* Expensive." Working Paper No. 86, Rock Center for Corporate Governance at Stanford University.

Aggarwal, Raj, Pamela Drake, Adam Kobor, and Gregory Noronha. 2011. "Capital Structure." Charlottesville, VA: CFA Institute.

Angel, James, Lawrence Harris, and Chester Spatt. 2011. "Equity Trading in the 21st Century." *Quarterly Journal of Finance*, Vol. 1: 1–53.

Appleyard, Dennis, Alfred Field, and Steven Cobb. 2010. *International Economics*, 7th edition. Boston: McGraw-Hill/Irwin.

Ariyoshi, Akira, Karl Habermeier, Bernard Laurens, Inci Otker-Robe, Jorge Iván Canales-Kriljenko, and Andrei Kirilenko. 2000. "Capital Controls: Country Experiences with Their Use and Liberalization." IMF Occasional Paper 190, Washington, DC (17 May).

Armantier, Olivier, Eric Ghysels, Asani Sarkar, and Jeffrey Shrader. 2011. "Stigma in Financial Markets: Evidence from Liquidity Auctions and Discount Window Borrowing during the Crisis." Federal Reserve Bank of New York Staff Reports, No. 483.

Ashauer, David. 1990. "Why Is Infrastructure Important?" In *Is There a Shortfall in Public Capital Investment?* Alicia Munnell, ed. Federal Reserve Bank of Boston Conference Series No. 34.

Bank for International Settlements (BIS). 2010. "Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity in 2010." www.bis.org.

Banker, R. D., I. Khosla, and K. K. Sinha. 1998. "Quality and Competition." *Management Science*, Vol. 44, No. 9: 1179–1192.

Bernanke, Ben S. 2009. "The Supervisory Capital Assessment Program." Atlanta Federal Reserve Bank Financial Markets Conference speech (11 May).

Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott. 2010. "Intrafirm Trade and Product Contractibility." *American Economic Review*, Vol. 100, No. 2 (May): 444–448.

Bureau of Labor Statistics. "Textile, Textile Product, and Apparel Manufacturing." In *Career Guide to Industries: 2010–11 Edition*.

Burns, Wesley Clair, and Arthur F. Mitchell. 1946. *Measuring Business Cycles*. Cambridge, MA: National Bureau of Economic Research.

Carson, John W. 2011. "Self-Regulation in Securities Markets." Working Paper No. 5542, World Bank Financial Sector Policy Group. http://ssrn.com/abstract=1747445.

Case, K., J. Quigley, and R. Shiller. 2005. "Comparing Wealth Effects: The Stock Market versus the Housing Market." *Advances in Macroeconomics*, Vol. 5, No. 1.

Chamberlin, Edward H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Christensen, Hans, Luzi Hail, and Christian Leuz. 2011. "Capital-Market Effects of Securities Regulation: The Role of Prior Regulation, Implementation and Enforcement." NBER Working Paper 16737 (October).

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy*, Vol. 113, No. 1: 1–45.

Coe, David T., and Elhanan Helpman. 1995. "International R&D Spillovers." *European Economic Review*, Vol. 39, No. 5 (May): 859–887.

Collier, Paul, and Stephen A. O'Connell. 2007. "Opportunities and Choices." In *The Political Economy of Economic Growth in Africa, 1960–2000*, Vol. 1. Benno J. Ndulu, Stephen A. O'Connell, Robert H. Bates, Paul Collier, and Charles C. Soludo, eds. Cambridge: Cambridge University Press.

Curcio, Riccardo, and Charles A. E. Goodhart. 1992. "Chartism: A Controlled Experiment." *Journal of International Securities Markets*, Vol. 7: 173–186.

Denison, Edward. 1985. *Trends in American Growth*. Washington, DC: Brookings Institution.

Dornbusch, Rudiger. 1976. "Expectations and Exchange Rate Dynamics." *Journal of Political Economy*, Vol. 84: 1161–1176.

Dorsey, Pat. 2004. *The Five Rules for Successful Stock Investing: Morningstar's Guide to Building Wealth and Winning in the Market*. Hoboken, NJ: John Wiley & Sons.

Espey, Molly. 1996. "Explaining the Variation in Elasticity Estimates of Gasoline Demand in the United States: A Meta-Analysis." *Energy Journal*, Vol. 17, No. 3: 49–60.

Feenstra, Robert C., and Alan M. Taylor. 2008. *International Economics*. New York: Worth Publishers.

Fleming, J. Marcus. 1962. "Domestic Financial Policies under Fixed and Floating Exchange Rates." *IMF Staff Papers*, Vol. 9: 319–379.

Friedman, Milton. 1953. "The Monetarist Theory of Flexible Exchange Rate Systems." In *Essays in Positive Economics*. Chicago: University of Chicago Press.

——— 1968. "The Role of Monetary Policy." *American Economic Review*, Vol. 58, No. 1: 1–17.

Friedman, Thomas L. 2006. *The World Is Flat: A Brief History of the Twenty-First Century*. New York: Farrar, Straus & Giroux.

Fudenberg, Drew, and Jean Tirole. 1984. "The Fat Cat Effect, the Puppy Dog Ploy and the Lean and Hungry Look." *American Economic Review*, Vol. 74, No. 2: 361–366.

Funke, N. 2004. "Is There a Stock Market Wealth Effect in Emerging Markets?" International Monetary Fund (March).

Gerber, James. 2010. *International Economics*, 5th edition. New York: Prentice Hall.

Gomez-Ibanez, Jose A. 2003. *Regulating Infrastructure: Monopoly, Contracts, and Discretion*. Cambridge, MA: Harvard University Press.

Goodhart, Charles A. E. 1989. "The Conduct of Monetary Policy." *Economic Journal*, Vol. 99, No. 396: 293–346.

Gray, Simon, and Nick Talbot. 2006. *Monetary Operations.* London: Bank of England. www.bankofengland.co.uk/education/ccbs/handbooks/ccbshb24.htm.

Greenspan, Alan. 2005. "Remarks on Central Banking." Speech given at the annual Kansas City Federal Reserve symposium in Jackson Hole, WY. Available online at www.federalreserve.gov/boarddocs/speeches/2005/20050826/default.htm.

Heston, Alan, Robert Summers, and Bettina Aten. 2009. Penn World Table Version 6.3. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania (August).

Hill, Charles W. L. 2007. *International Business: Competing in the Global Marketplace*, 6th edition. Boston: Irwin/McGraw-Hill.

Hong Kong Monetary Authority (HKMA). 2005. *HKMA Background Brief No. 1: Hong Kong's Linked Exchange Rate System*, 2nd edition (November). www.info.gov.hk.

International Monetary Fund. 1998. *IMF World Economic Outlook* (May 1998), Washington, DC: International Monetary Fund.

——— 2006. "De Facto Classification of Exchange Rate Regimes and Monetary Policy Framework." www.imf.org.

——— 2007. "Exchange Rates and the Adjustment of External Imbalances." *IMF World Economic Outlook* (April 2007), Ch. 3, 81–120.
——— 2008. *Globalization: A Brief Overview*. Issues Brief, International Monetary Fund (May).
——— 2009. "The Case for Global Fiscal Stimulus" (March). www.imf.org/external/pubs/ft/spn/2009/spn0903.pdf.
——— 2010a. *Balance of Payments and International Investment Position Manual*, 6th edition. Washington, DC: International Monetary Fund.
——— 2010b. "Currency Composition of Official Foreign Exchange Reserves (COFER)" report. www.imf.org.
——— 2010c. *World Economic Outlook: April 2010.* Washington, DC: International Monetary Fund.
——— 2011a. "The IMF at a Glance." International Monetary Fund (February). www.imf.org/external/np/exr/facts/glance.htm.
——— 2011b. "Statistical Appendix." In *World Economic Outlook* (September). Washington, DC: International Monetary Fund.
Isard, Peter, Hamid Faruqee, G. Russell Kincaid, and Martin Fetherston. 2001. "Methodology for Current Account and Exchange Rate Assessments." IMF Occasional Paper #209.
Jorgenson, Dale. 1966. "Technology in Growth Theory." Technology and Growth, Federal Reserve Bank of Boston Conference Series.
Jorgenson, Dale. 2000. "Raising the Speed Limit: U.S. Economic Growth in the Information Age." Brooking Papers on Economic Activity.
Kaul, Aditya, Vikas Mehrotra, and Randall Morck. 2000. "Demand Curves for Stocks Do Slope Down: New Evidence from an Index Weights Adjustment." *Journal of Finance*, Vol. 55 (2 April): 893–912.
Kawai, Masahiro, and Shinji Takagi. 2003. "Rethinking Capital Controls: The Malaysian Experience." PRI Discussion Paper Series No. 03A-05, Policy Research Institute, Ministry of Finance Japan, Tokyo (May).
Kelly, Anthony. 2003. *Decision Making Using Game Theory: An Introduction for Managers*. Cambridge: Cambridge University Press.
Klitgard, Thomas, and Laura Weir. 2004. "Exchange Rate Changes and Net Positions of Speculators in the Futures Market." *Economic Policy Review*. Federal Reserve Bank of New York (May).
Krugman, Paul R. 1989. "Industrial Organization and International Trade." In *Handbook of Industrial Organization*, Vol. 2. Richard Schmalensee and Robert Willig, eds. Amsterdam: Elsevier, B.V.
Levine, R. 2005. "Finance and Growth: Theory and Evidence." In *Handbook of Economic Growth*. Philippe Aghion and Steven Durlauf, eds. Amsterdam: Elsevier, B.V.
Magud, Nichols E., Carmen M. Reinhart, and Kenneth S. Rogoff. 2011. "Capital Controls: Myth and Reality—A Portfolio Balance Approach." Peterson Institute for International Economics, Working Paper No. 11-7.
Mankiw, N. Gregory. 1989. "Real Business Cycles: A New Keynesian Perspective." *Journal of Economic Perspectives*, Vol. 3, No. 3: 79–90.
Mankiw, Gregory. 1995. "The Growth of Nations." Brookings Papers on Economic Activity.
McCloskey, D. 1982. *The Applied Theory of Price*, 2nd edition. New York: Macmillan.
McCulley, Paul. 2010. "The Shadow Banking System and Hyman Minsky's Economic Journey." In *Insights into the Global Financial Crisis*. Charlottesville, VA: Research Foundation of CFA Institute.
McGuigan, James R., R. Charles Moyer, and Frederick H. Harris. 2008. *Managerial Economics: Applications, Strategy and Tactics*, 11th edition. Mason, OH: Thomson South-Western.
Meier, Gerald M. 1998. *The International Environment of Business: Competition and Governance in the Global Economy*. New York: Oxford University Press.
Mogford, Caroline, and Darren Pain. 2006. "The Information Content of Aggregate Data on Financial Futures Positions." *Bank of England, Quarterly Bulletin* (Spring 2006).
Mundell, Robert A. 1962. "The Appropriate Use of Monetary and Fiscal Policy for Internal and External Stability." *IMF Staff Papers*, Vol. 9: 70–79.
——— 1963. "Capital Mobility and Stabilization Policy under Fixed and Flexible Exchange Rates." *Canadian Journal of Economics and Political Science*, Vol. 29: 475–485.

Nicholson, Walter, and Christopher M. Snyder. 2008. *Microeconomic Theory: Basic Principles and Extensions*, 10th edition. Mason, OH: Thomson South-Western.

OECD. 2003. *The Sources of Economic Growth in the OECD Countries*. Paris: Organization for Economic Cooperation and Development.

Plosser, Charles I. 1989. "Understanding Real Business Cycles." *Journal of Economic Perspectives*, Vol. 3, No. 3: 51–77.

Porter, Michael E. 1980. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.

Pukthuanthong-Le, Kuntara, and Lee R. Thomas, III. 2008. "Weak-Form Efficiency in Currency Markets." *Financial Analysts Journal*, Vol. 64, No. 3: 31–52.

Reinhardt, Carmen, and Kenneth Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, NJ: Princeton University Press.

Roberts, Mark, and Uwe Deichmann. 2008. "Regional Spillover Estimation." Background paper for the *World Development Report 2009: Reshaping Economic Geography*, World Bank.

Roger, Scott. 2010. "Inflation Targeting Turns 20." *Finance and Development*, Vol. 47, No. 1 (March): 46–49.

Romano, Roberta. 2005. "The Sarbanes-Oxley Act and the Making of Quack Corporate Governance." *Yale Law Journal*, Vol. 114, No. 7 (May): 1521–1611.

Romer, David. 2005. *Advanced Macroeconomics*, 2nd edition. Columbus, OH: McGraw-Hill.

Romer, Paul. 1986. "Increasing Returns and Long-Run Growth." *Journal of Political Economy*, Vol. 94, No. 5 (October): 1002–1037.

Rosenberg, Michael R. 1996. *Currency Forecasting: A Guide to Fundamental and Technical Models of Exchange Rate Determination*. Chicago: Irwin Professional Publishing.

——— 2002. *Deutsche Bank Guide to Exchange Rate Determination*. Chicago: Irwin Professional Publishing.

——— 2003. *Exchange Rate Determination: Models and Strategies for Exchange Rate Forecasting*. New York: McGraw-Hill.

Salvatore, Dominick. 2010. *Introduction to International Economics*, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Schumpeter, Joseph A. 1942. *Capitalism, Socialism and Democracy*. New York: HarperCollins.

Siegel, Lawrence. 2010. *Insights into the Global Financial Crisis*. Charlottesville, VA: Research Foundation of CFA Institute.

Solow, Robert. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics*, Vol. 39, No. 3 (August): 312–320.

Spatt, Chester. 2009. "Regulatory Conflict: Market Integrity vs. Financial Stability." *University of Pittsburgh Law Review*, Vol. 71, No. 3 (Winter): 625–639.

——— 2011. "Measurement and Policy Formulation." Invited Lecture to the Society for Financial Econometrics meeting at the University of Chicago (June).

Stewart, Scott, Christopher Piros, and Jeffrey Heisler. 2011. *Running Money: Professional Portfolio Management*. New York: McGraw-Hill/Irwin.

Stigler, George J. 1971. "The Economic Theory of Regulation." *Bell Journal of Economics*, Vol. 2, No. 1 (Spring): 3–21.

Suttle, Philip, Robin Koepke, Kristina Morkunaite, and Emre Tiftik. 2011. "Capital Flows to Emerging Market Economies." IIF Research Note (September 25). Institute of International Finance.

Taylor, John B. 1993. "Discretion versus Policy Rules in Practice." Carnegie-Rochester Conference Series on Public Policy, Vol. 39: 195–214.

Truman, Edwin. 2003. *Inflation Targeting in the World Economy*. Washington, DC: Institute for International Economics.

United Nations. 2002. *World Investment Report 2002: Transnational Corporations and Export Competitiveness*. New York: United Nations Conference on Trade and Development (UNCTAD).

von Stackelberg, Heinrich. 1952. *The Theory of the Market Economy*. New York: Oxford University Press.
Woodford, Michael. 2009. "Convergence in Macroeconomics: Elements of the New Synthesis." *American Economic Journal: Macroeconomics*, Vol. 1, No. 1: 267–279.
World Bank. 2009. *World Development Report 2009: Reshaping Economic Geography*. Washington, DC: World Bank.
World Trade Organization. 2008. *World Trade Report 2008: Trade in a Globalizing World*. Geneva: World Trade Organization.

ABOUT THE EDITORS

Christopher D. Piros, PhD, CFA, is the Managing Director of Investment Strategy and Chairman of the Investment Policy Committee at Hawthorn, a member of the PNC Financial Services Group, Inc., dedicated to serving the needs of individuals and families with investable assets in excess of $20 million.

He joined PNC from CFA Institute, where he served on the team responsible for the curriculum underlying the Chartered Financial Analyst designation. Previously, he was Director of Investment Strategy and Portfolio Management at Prudential Investments LLC, the wealth management services arm of Prudential Financial, where he established and led the unit's discretionary portfolio management activities. He and his team also formulated the investment strategy advice disseminated through the firm's wealth management platform, developed its proprietary asset allocation methodology, and counseled Prudential's investment boards on the outlook for capital markets and the global economy. Earlier he was a global fixed-income portfolio manager and head of fixed-income quantitative analysis at MFS Investment Management.

Dr. Piros earned his PhD in economics at Harvard University and began his career on the finance faculty of Duke University's Fuqua School of Business. He currently teaches in the MS in Investment Management programs at both Boston University and Reykjavik University and is on the advisory board of the program at BU. He is a coauthor of *Running Money: Professional Portfolio Management* (McGraw-Hill, 2011).

Chris is a CFA charterholder and a member of CFA Institute, the New York Society of Securities Analysts, and the Chicago Quantitative Alliance (CQA).

Jerald E. Pinto, PhD, CFA, is Director of Curriculum Projects in the Education Division of CFA Institute. Before coming to CFA Institute in 2002, he consulted to corporations, foundations, and partnerships in investment planning, portfolio analysis, and quantitative analysis. He also worked in the investment and banking industries in New York City from the late 1970s on and taught investments at New York University's Stern School of Business. He was secretary-treasurer and responsible for the investments of Friends of Laurentian University, Inc., from 1986 to 2002. He served CFA Institute in a number of volunteer capacities before joining the staff and was a judge for the CFA Virginia Investment Research Challenge in 2008 and 2009. He is a coauthor of *Quantitative Investment Analysis* (2007) and a coeditor of and contributor of chapters to the third edition of *Managing Investment Portfolios: A Dynamic Process* (2007). He holds an MBA from Baruch College and a PhD in finance from the Stern School and is a member of CFA Virginia.

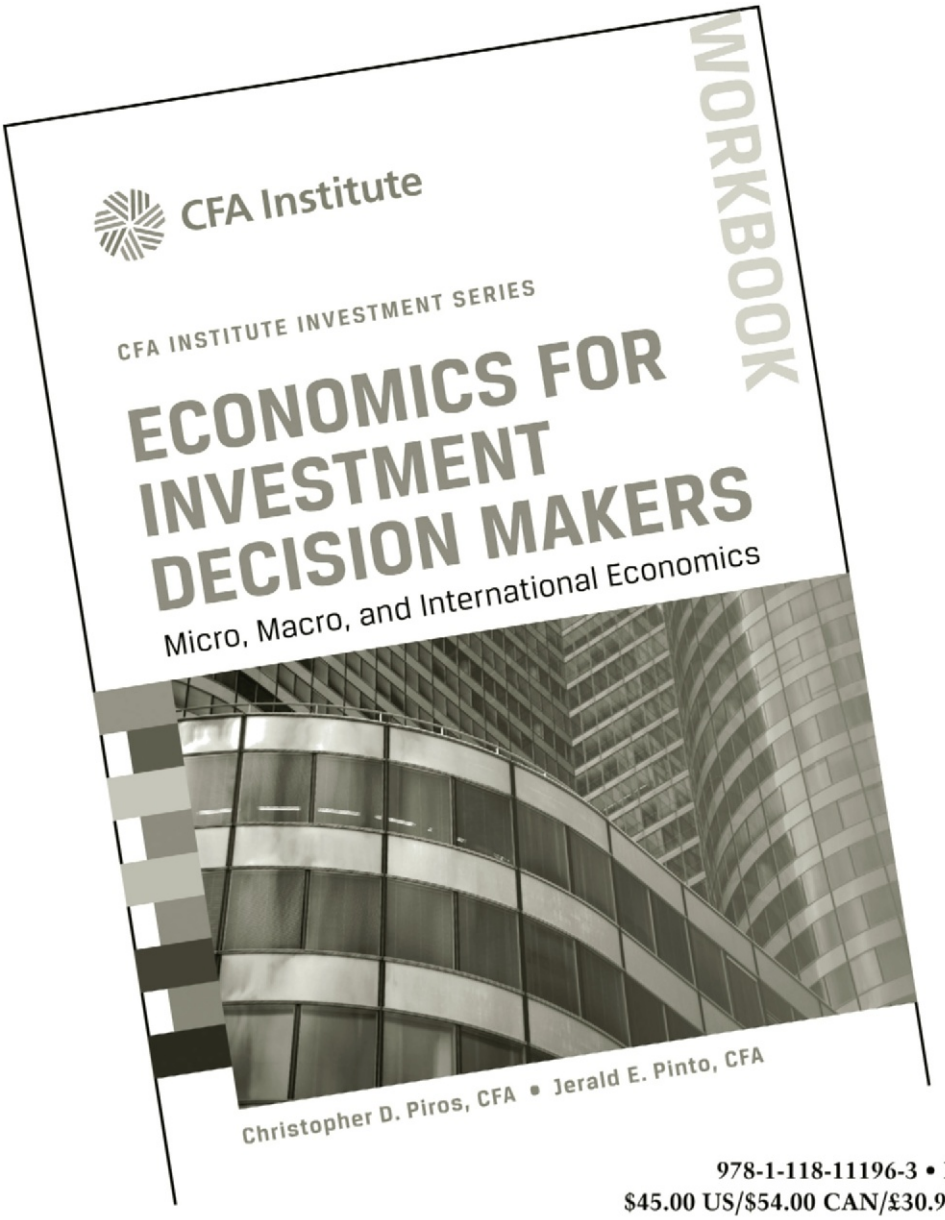
ABOUT THE CFA PROGRAM

The Chartered Financial Analyst (CFA) designation is a globally recognized standard of excellence for measuring the competence and integrity of investment professionals. To earn the CFA charter, candidates must successfully pass through the CFA Program, a global graduate-level self-study program that combines a broad curriculum with professional conduct requirements as preparation for a wide range of investment specialties.

Anchored by a practice-based curriculum, the CFA Program is focused on the knowledge identified by professionals as essential to the investment decision-making process. This body of knowledge maintains current relevance through a regular, extensive survey of practicing CFA charterholders across the globe. The curriculum covers 10 general topic areas, ranging from equity and fixed-income analysis to portfolio management to corporate finance—all with a heavy emphasis on the application of ethics in professional practice. Known for its rigor and breadth, the CFA Program curriculum highlights principles common to every market so that professionals who earn the CFA designation have a thoroughly global investment perspective and a profound understanding of the global marketplace.

www.cfainstitute.org

INDEX



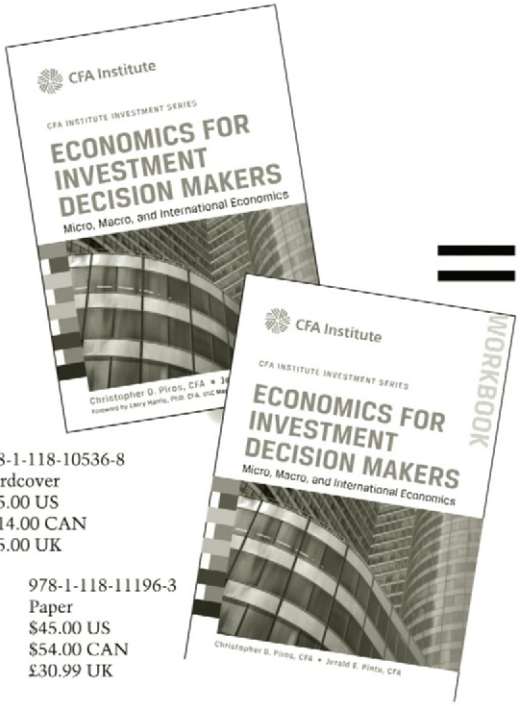
978-1-118-11196-3 • Paper
$45.00 US/$54.00 CAN/£30.99 UK

The **essential companion workbook** to the tool of choice for investment decision makers.

Available at wiley.com, cfainstitute.org, and wherever books are sold.

WILEY | CFA Institute

CFA Institute + Wiley = Success

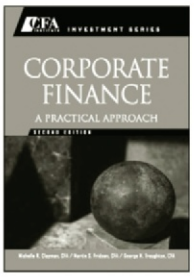


978-1-118-10536-8
Hardcover
$95.00 US
$114.00 CAN
£65.00 UK

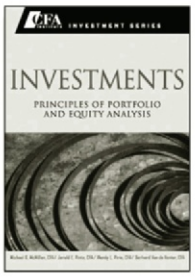
978-1-118-11196-3
Paper
$45.00 US
$54.00 CAN
£30.99 UK

John Wiley & Sons and CFA Institute are proud to present the *CFA Institute Investment Series* geared specifically for industry professionals and graduate-level students. This cutting-edge series focuses on the most important topics in the finance industry. The authors of these books are themselves leading industry professionals and academics who bring their wealth of knowledge and expertise to you.

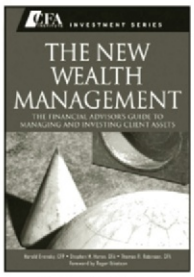
The series provides clear, practitioner-driven coverage of the knowledge and skills critical to investment analysts, portfolio managers, and financial advisors.



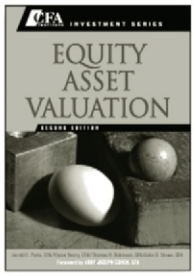
978-1-118-10537-5
$95.00 US
$114.00 CAN/£65.00 UK



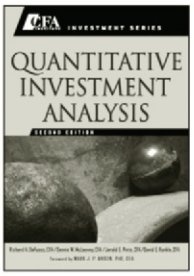
978-0-470-91580-6
$95.00 US
$114.00 CAN/£65.00 UK



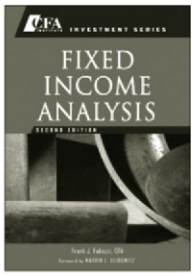
978-0-470-62400-5
$95.00 US
$114.00 CAN/£65.00 UK



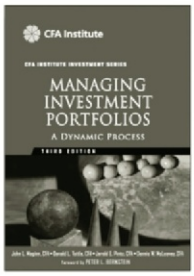
978-0-470-57143-9
$95.00 US
$114.00 CAN/£65.00 UK



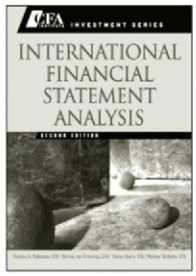
978-0-470-05220-4
$95.00 US
$104.00 CAN/£65.00 UK



978-0-470-05221-1
$95.00 US
$104.99 CAN/£65.00 UK



978-0-470-08014-6
$100.00 US
$119.99 CAN/£70.00 UK



978-0-470-91662-9
$95.00 US
$114.00 CAN/£65.00 UK

Get these titles and companion Workbooks at wiley.com or cfainstitute.org.
Available in print and e-book format.

WILEY | CFA Institute