FOURTH EDITION

THE ECONOMICS OF HEALTH *Reconsidered*

THOMAS RICE *and* LYNN UNRUH

THE ECONOMICS OF HEALTH *Reconsidered*

FOURTH EDITION

THE ECONOMICS OF HEALTH *Reconsidered*

THOMAS RICE *and* LYNN UNRUH

AUPHA

Health Administration Press, Chicago, Illinois

Association of University Programs in Health Administration, Arlington, Virginia

Your board, staff, or clients may also benefit from this book's insight. For more information on quantity discounts, contact the Health Administration Press Marketing Manager at (312) 424-9470.

This publication is intended to provide accurate and authoritative information in regard to the subject matter covered. It is sold, or otherwise provided, with the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The statements and opinions contained in this book are strictly those of the author and do not represent the official positions of the American College of Healthcare Executives, the Foundation of the American College of Healthcare Executives, or the Association of University Programs in Health Administration.

20 19 18 17 16 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Rice, Thomas H., author.
The economics of health reconsidered / Thomas Rice and Lynn Unruh. — Fourth edition.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-56793-723-7 (alk. paper)
I. Unruh, Lynn, author. II. Association of University Programs in Health Administration, issuing body. III. Title.
[DNLM: 1. Economics, Medical—United States. 2. Cost-Benefit Analysis—United States. 3. Health Expenditures—United States. 4. Health Services Administration—economics—United States. 5. Models, Economic—United States. W 74 AA1]
RA410
338.4'33621—dc23
2015000968

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. ♾™

Acquisitions editor: Tulie O'Connor; Project manager: Andrew Baumann; Cover designer: Marisa Jackson; Layout: PerfecType

Found an error or a typo? We want to know! Please e-mail it to hapbooks@ache.org, and put "Book Error" in the subject line.

Health Administration Press
A division of the Foundation of the American College of Healthcare Executives
One North Franklin Street, Suite 1700
Chicago, IL 60606-3529
(312) 424-2800

Association of University Programs in Health Administration
2000 North 14th Street
Suite 780
Arlington, VA 22201
(703) 894-0940

In memory of Gavin Mooney

BRIEF CONTENTS

DETAILED CONTENTS

PREFACE TO THE FOURTH EDITION

We are pleased to have the opportunity to provide a fourth edition of *The Economics of Health Reconsidered*. Like the third edition, this one has been designed to be used as a stand-alone textbook for graduate and advanced undergraduate courses in health economics, or in conjunction with key journal articles in the field. Instructors should note that there is an accompanying password-protected instructor's manual that provides a list of concepts, discussion questions, and additional readings for each of the main chapters in the book. Moreover, for the first time, an extensive set of PowerPoint slides is provided for each chapter. For access to these resources, e-mail hapbooks@ache.org.

All chapters, figures, and tables have been thoroughly updated in this edition. We have added two new chapters to the book:

- Chapter 10 provides a close examination of healthcare expenditures, both in the United States and in other high-income countries. It includes information on causes of high expenditures, their magnitude and growth, and different policies that have been used or proposed to help control them.
- Chapter 11 reviews the topic of economic evaluation in healthcare. It describes the types of economic evaluation in use in healthcare, provides some training in how to conduct an economic evaluation, and discusses issues related to the use of economic evaluations.

This book has had many updates in the 17 years since the first edition was published, but its basic theme has remained the same: Despite assertions to the contrary, neither economic theory nor evidence shows that reliance on market forces leads to superior outcomes in healthcare systems. Government has a crucial role to play in making the sector not only more equitable but more efficient as well.

ACKNOWLEDGMENTS

We would like to express our deep appreciation to a number of people who provided comments on the research itself or on the individual chapters, not only in this edition but in the previous three editions as well: Henry Aaron, Ronald Andersen, Gerard Anderson, William Comanor, Janet Cummings, Katherine Desmond, Brian Elbel, Robert Evans, Rashi Fein, Paul Feldstein, Susan Haber, Yaniv Hanoch, Diana Hilberman, Miriam Laugesen, Donald Light, Harold Luft, David Mechanic, Glenn Melnick, Gavin Mooney, Jack Needleman, Joseph Newhouse, Mark Peterson, Uwe Reinhardt, John Roemer, Sally Stearns, Greg Stoddart, Deborah Stone, Ewout van Ginneken, Pete Welch, Joseph White, and Miriam Wiley. We would also like to thank Jimmy Alex, Mona Naroozi, and Ashley Rutherford for their excellent research assistance on the current edition.

All conclusions, and any errors, are entirely our own and are not the responsibility of any of the reviewers.

PART

I

INTRODUCTION

As the book's title indicates, the economics of health needs to be reconsidered. While health economists recognize the need for government involvement in the marketplace, they still tend to advocate reliance on market forces as the solution for most of the ills faced by healthcare systems. This book questions the wisdom of this mind-set, using theory and empirical evidence.

To understand the advisability of alternative reform methods, one must first understand the traditional competitive model. After providing a context for the book in Chapter 1—where we make the case that health economic theory needs to be reconsidered—we present a detailed summary of microeconomic theory in Chapter 2. That chapter explains the key tools of the trade: demand, supply, competition, monopoly, and social welfare. The remainder of the book examines the assumptions underlying the competitive model, whether they are met in the healthcare realm, and the advisability of alternative ways of reforming healthcare markets.

Chapter 3 lists 14 assumptions that need to be met to ensure socially optimal results from the use of market forces. Later chapters examine whether each of these assumptions is met; we provide evidence that they are not. This discussion does not mean government intervention is necessarily superior, however. One must evaluate empirically where markets succeed and fail. Indeed, just as markets fail, so can government. Of course, all countries use markets and governments in varying degrees, so it is not an either/or choice but a matter of emphasizing the appropriate policy tools in a specific proposed reform. One of the book's key points is that because so few of the assumptions of competitive markets are met in healthcare, one cannot presume that pro-competitive policies will be superior.

CHAPTER

1

WHY SHOULD THE ECONOMICS OF HEALTH BE RECONSIDERED?

1.1 Context

Recent years have seen a surge of interest in reforming the organization and delivery of health systems by replacing government regulation with reliance on market forces. Although much of the impetus has come from the United States, the phenomenon is worldwide. Spurred by ever-increasing costs coupled with competing priorities such as education, welfare, and, more recently, environmental concerns, analysts and policymakers have embraced the competitive market as the means of choice for reforming medical care systems. To a great extent, this belief stems from economic theory, which purports to show the superiority of markets over strong government involvement.

The United States is a case in point. Two recent examples reflect the way in which health insurance has been extended to segments of the population. In 2006, the Medicare program, which services older and disabled Americans, was expanded to include prescription drugs. This expansion was implemented by having the new benefits provided by competing, private insurance companies. Similarly, when the Affordable Care Act was being debated in 2009 and 2010, President Obama called for a public insurance option as an alternative to compete against private insurers, but this was ultimately rejected such that coverage for previously uninsured individuals can only be provided by the private sector. Other countries have followed a similar path. Most notable is the Dutch healthcare system, which in 2006 implemented major reforms to its universal healthcare system by embracing the notion of competing private insurers. The perceived success of this increasingly competitive marketplace in healthcare sectors is part of a broader trend in the United States, in which markets are viewed as efficient and government is viewed as inefficient. As Robert Kuttner (1997) wrote, "America . . . is in one of its cyclical romances with a utopian view of laissez-faire." The relevance of this statement persists almost two decades later because the cycle has not yet ended. We do not imply, either in the health sector or in the economy as a whole, that policymakers have eschewed government involvement. Our concern is that healthcare markets are moving in this direction and that economic theory is used—inappropriately, we will argue—in support of market-based health policies.

The intellectual case for relying on markets in health is based in part on the writings of Alain Enthoven (1978a, 1978b, 1988, 2003; Enthoven and Kronick 1989a, 1989b), who advocates reliance on consumer choice and competition to improve the efficiency of healthcare markets. Nevertheless, he still believes that government has two key roles: ensuring that competition is based on price rather than selection of the healthiest patients and providing subsidies to low-income persons.

The corollary to this viewpoint is that government should *confine* itself to these two roles. Health services policy should be based on competition, with government ensuring that markets operate fairly and helping disadvantaged people. A careful review of economic theory as applied to health, however, does not permit government such a limited role.

This book contends that one of the main justifications for the superiority of market-based systems stems from a misapplication of economic theory to health. As we will show, this application is based on a large set of assumptions that are not met and cannot be met in the healthcare sector. This contention does not mean that competitive approaches in this key sector of the economy are inappropriate; rather, their efficacy depends on the policy being considered and the environment in which it is to be implemented. Stated more colloquially, it works well in some instances but not in others. There is, however, no reason to believe market-based systems will operate more efficiently or provide a higher level of social welfare than alternative systems based on governmental financing and regulation. This argument is further bolstered by the deviation of many other developed countries from market-based health systems.

Although economists know that claims about the superiority of competitive approaches are based on fulfillment of assumptions, the healthcare literature rarely mentions the large number of such assumptions or their importance. One should not put undue blame on health economists, however; this problem pervades the entire economic discipline. In this regard, Lester Thurow (1983) has written that "every economist knows the dozens of restrictive assumptions . . . that are necessary to 'prove' that a free market is the best possible economic game, but they tend to be forgotten in the play of events." Chapter 3 provides our list of these assumptions, and in subsequent chapters we show their implications in the fields of health economics and health policy.

The book thus centers on the description, analysis, and application of the assumptions on which the superiority of competition is based—and in particular, what happens in markets for health services if they are not met.

1.2 Purpose of the Book

The purpose of this book is to reconsider the economics of health. It does so by examining the assumptions on which the superiority of competitive approaches is based and how failure to meet those assumptions affects health policy choices.

Although each chapter provides applications, the book is also about theory—its use and its misuse. The book will attempt to show that economic theory does not support the belief that competition in the health services sector will necessarily lead to superior social outcomes.

If economic theory does not demonstrate the superiority of market forces in health, questions must be answered empirically. To a large extent, that is exactly what health economists and health services researchers are trying to do. We have few reservations about the kinds of research studies being conducted. Our concern is that the work will suffer if researchers approach it with preconceived notions of what the results ought to be.

Some readers will be disappointed to see that although the book critiques the competitive model, it does not explicitly offer a theoretical alternative. It does, however, compare the health systems of countries that use varying ratios of government and markets. Ultimately, readers must draw their own conclusions about the most desirable system using theory and the extant empirical literature. We hope this book can help them do so.

The fourth edition of this book adds two new chapters. Chapter 10 presents data and analyzes healthcare expenditures and trends. It includes a discussion of measuring expenditures, presents data from the United States and other countries on expenditure trends, and analyzes drivers and methods used to control expenditures in different countries. Chapter 11 is more methodological, focusing on different ways of conducting economic evaluations in healthcare. It includes cost–benefit, cost-effectiveness, and cost–utility analysis as well as comparative effectiveness analysis.

The book is also addressed to noneconomics professions. Because students and practitioners in these disciplines obviously tend to be less schooled in the details of economic analysis, they often have to take health economists at their word when the latter speak about the policy implications of economic analysis in general, and the superiority of markets in particular. (In this regard, Joan Robinson has been quoted as advising, “Study economics to avoid being deceived by economists” [Kuttner 1984].) We hope this book will help put those in disciplines other than economics on a level playing field when it comes to discussions of health policy.

1.3 Outline of the Book

The book is divided into 12 chapters and a conclusion. Chapter 2 covers nearly all of the major topics a course in microeconomic theory would cover. A few remaining topics (e.g., externalities, labor economics) are discussed later in the book. Those who are already familiar with intermediate microeconomic theory can proceed directly to the other chapters. Others may want to refer to Chapter 2 when reading the subsequent material.

Chapter 3 provides a list of the assumptions on which the superiority of market competition is based, as well as an overview of the role of government. We critique those assumptions in the chapters that follow. Chapter 4 focuses on the theory of demand, and Chapter 5 applies the theory of demand to health insurance and particular health services. Chapters 6 through 8 focus on supply: issues of competition and market power in healthcare supply and demand, for-profit medicine, and workforce issues, respectively. Chapter 9 explores equity and redistribution, a topic of tremendous importance to policy but one that has received insufficient attention from health economists. As noted, Chapters 10 and 11 are new to the book and focus, respectively, on healthcare expenditures and on how to conduct economic evaluations. Chapter 12 discusses different ways developed countries can organize, and have organized, their healthcare systems, and it includes cross-national empirical evidence on outcomes and costs and tentative lessons from this evidence. The conclusion offers some final thoughts concerning the role of competition in the healthcare sector.

CHAPTER

2

THE TRADITIONAL COMPETITIVE MODEL

The field of microeconomics is devoted to the study of competition—mainly its virtues, but also some of its pitfalls. Although many of the techniques economists use are fairly new, the emphasis on competition dates back more than 200 years, to the writings of Adam Smith (1776 [1994]). Smith believed that people driven by their own economic interest in the marketplace are guided by an "invisible hand" to act in the manner that most benefits society at large. The concept that societal outcomes are optimal when individuals and firms act in what one might view as a completely selfish manner is a key insight of economic theory. As we will explain later in this chapter, the word *optimal* has a specific economic meaning that differs from the word's common definition.

The notion of competition is intuitively appealing. In a competitive market, people are allowed but not compelled to trade their wealth, including their labor, if they find it beneficial to do so. Theoretically, when everyone stops trading because there is nothing more to gain, the market is in *equilibrium*. Such an outcome is desirable in two senses: (1) People are making their own choices, and (2) by not engaging in any more trades, people *reveal themselves* to be as satisfied as possible with their economic lot, given the resources with which they began. Analogously, firms can enter and exit the market at will and produce as much or as little as they wish. To beat the competition, however, they will endeavor to produce only what people demand, using the fewest possible resources to obtain the highest profits. This action leaves more resources available to fulfill demand for other products and services.

This chapter outlines the economic theory of competition and what competition can and cannot achieve. The chapter is divided into five subsections: utility and demand; production, costs, and supply; equilibrium in a competitive market; equilibrium for a monopolist; and the economy as a whole. A few other microeconomic issues, particularly externalities (Chapter 5) and labor supply (Chapter 8) are discussed in subsequent chapters where the topics naturally arise.

One must understand the basics of microeconomic theory to appreciate the book's critiques of its application to health. Readers who are familiar with the standard theory can proceed immediately to Chapter 3, while those seeking more detail may wish to consult a microeconomics textbook.

2.1 Utility and Demand

Utility

We begin with the notion of *utility*. Perhaps the easiest way to think about this term is through some synonyms, such as *happiness, satisfaction,* or even *physical and mental well-being*. One of the key concepts of microeconomic theory is that consumers attempt to maximize their utility.

The utility obtained from the consumption of one more unit of a good is its *marginal utility*. Economists generally believe that the marginal utility a person receives from a particular good declines as that person obtains more of the good. For example, having one automobile might give you a lot of utility, and having a second might give you more—but not as much as the first one did. This concept is called *diminishing marginal utility*.

Some early theorists, collectively known as the *classical utilitarians*, believed that one could compute how well off an entire society was by adding up each person's utility. But to do this calculation, we would need to assign a quantitative value to the utility possessed by each person in a society. This conundrum naturally led to the question of whether it was possible to quantify such measures. A leading early advocate of classical utilitarianism, Jeremy Bentham, thought it was possible to measure utility through its manifestations of pleasure and pain. In the colorful language of the early nineteenth century, he wrote:

> Nature has pleased mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think: every effort we can make to throw off our subjection, will serve but to demonstrate and confirm it. (Bentham 1968)

Economics is rarely viewed as a left-wing social science—how could it be if it is based on an axiom of self-interest? But if, as the classical utilitarians claimed, everyone has the same capacity to experience pleasure and pain, and if we assume the existence of diminishing marginal utility, then social welfare is maximized when everyone has the same income. For example, suppose one person has $10,000 in income and another has $5,000. If an additional dollar is spent by the former person, it will bring less utility than if the same dollar were spent by the latter. Only if everyone has the same income will total welfare be maximized.

Although some radical utilitarians were comfortable with this concept, others—some of whom presumably would have lost a great deal through the equalization of incomes—were not. And it was not hard to poke holes in the

theory. Classical utilitarianism is based on two key assumptions: (1) Utility can be quantified, and (2) it is possible to add utilities across different individuals. That is, everyone has a common quantitative metric, whereby, say, three units of utility (or *utiles*) for you is equivalent to the same number of utiles for me.

Modern economists eschew these assumptions in favor of a milder form of utilitarian principles. For a century, microeconomics has proceeded under the *Pareto principle*, where a policy is considered desirable if it makes someone better off without making anyone worse off. But to go further—say, to advocate one public program or tax over another as better for society as a whole, when there will be winners and losers—requires explicit value judgments. Despite occasional claims to the contrary, economic theory almost never implies that one policy is better than another, because this would involve weighing the benefits that accrue to one group against the losses incurred by the other. The most theory can do is demonstrate the advantages and disadvantages of each alternative (Culyer 2012).

Indifference Curves

Recall that economic theory assumes that people seek to maximize their utility. Utility, the outcome, is on the left side of Equation 2.1, and the determinants of utility, an example of which is the goods and services people consume, are on the right side.

Let:

$$U = f(X, Y, Z, \ldots, n) \tag{2.1}$$

where U is a person's utility (f denotes function). There are n goods or services that a person consumes, three of which are labeled X, Y, and Z. The possession of this bundle of goods constitutes the person's utility level, U. We further assume that although there is diminishing marginal utility, consumers do not reach a *saturation point*, at which an additional unit of X, Y, or Z actually reduces their utility. In other words, people are happier when they have more things—an issue we examine in Chapters 4 and 5.

Another important and somewhat hidden assumption inherent in this theory is: People are affected only by the things they possess and are unaffected by what others have or by how their bundle of goods compares with those of others. This assumption only becomes apparent if we explicitly denote that we are dealing with only a representative individual, person i.

$$U_i = f(X_i, Y_i, Z_i, \ldots, n_i) \tag{2.2}$$

Clearly, this person's utility is only affected by what he has, not by what others have. We can represent an alternative scenario in which people are affected by both their own possessions and those of others by including another subscript for a representative other individual, *j*.

$$U_i = f(X_i, X_j, Y_i, Y_j, Z_i, Z_j, \ldots, n_i, n_j) \tag{2.3}$$

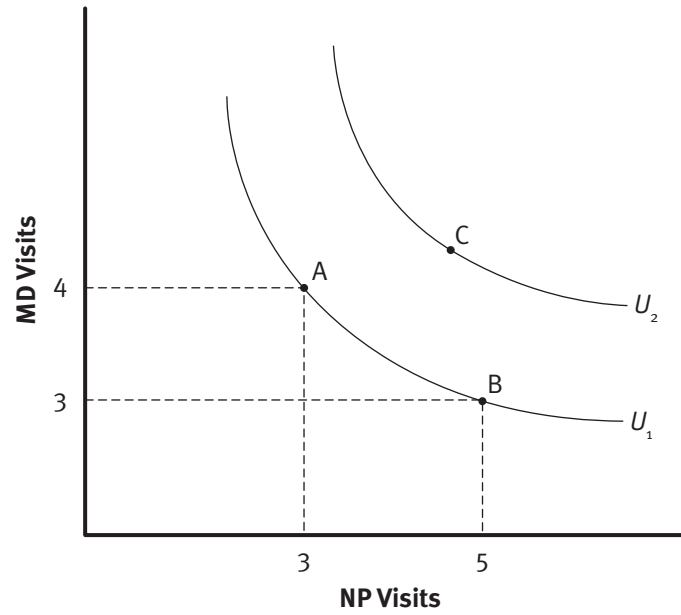
The implications of such *interdependent utilities* are examined further in Chapter 5.

The conventional theory assumes people seek to maximize their utility, which, as we noted, is determined by the bundle of goods and services they possess. To do so, they purchase their ideal bundle based on their desire or *taste* for the alternative goods and the prices of these alternatives, subject, of course, to how much income they have available to spend.

We will use graphs throughout this chapter, as they are helpful in illustrating these concepts. In doing so, however, we can show at most only two of the many goods and services people wish to have—one on each axis.

Exhibit 2.1 shows two *indifference curves*, which represent alternative combinations of two goods that result in the same level of utility. Imagine that a person spends all of his income on these two goods and there are no savings. While these concepts are abstractions, they make the theory easier to understand.

EXHIBIT 2.1 Consumer Indifference Curves



Throughout this chapter, we examine two competing goods: nurse practitioner (NP) visits and physician (MD) visits. The consumer is indifferent to all points on a particular curve because by definition all points bring equal levels of satisfaction. Three NP visits and four MD visits (point A) are equal in desirability to five NP visits and three MD visits (point B) on curve U_1. The person would be even happier to have more (e.g., point C on curve U_2), but that might involve spending more money than he has available.

Curve U_2 conveys a higher utility than U_1 because at each point, the person possesses more of both goods. In theory, a consumer has an infinite number of indifference curves, each corresponding to different combinations of quantities of the two goods.

The typical indifference curve has three characteristics. First, it tends to have a convex-to-the-origin shape because of diminishing marginal utility. Once a person has a great deal of one good and little of another, that person has to receive a lot more of the former to give up even a little bit of the latter. The slope of the indifference curve, which of course varies at each point if it is not a straight line, is called the *marginal rate of substitution*. The rate is equal to the ratio of the marginal utilities of the two goods.[1]

Second, indifference curves don't bend all the way back around. Stated more technically, they never exhibit a positive slope. A given quantity on the x- or y-axis corresponds to only one point on an indifference curve. The indifference curve implies that consumers do not reach a satiation point—they always get more utility from an additional unit of a good, no matter how much they already have.

Third, two indifference curves cannot intersect. If they did, then all points on *both* curves would confer the same amount of utility. If one were to draw such an exhibit, there would be points on the graph where having more of both goods would not bring higher utility, which violates the definition of the curve.

The Budget Constraint

The choice of how much of each type of visit to purchase depends not only on how much the person wants of each visit type but also on the price of each type. We can illustrate this concept using another graphical tool, the *budget constraint*. It is a line that shows how many of each type of visit the consumer can purchase with a given income, and it can be derived through Equation 2.4,

$$I = (P_M \times Q_M) + (P_{NP} \times Q_{NP}) \tag{2.4}$$

and solving for Q_M by rearranging the terms, as in Equation 2.5:

$$Q_M = I / P_M - [(P_{NP} / P_M) \times Q_{NP}], \tag{2.5}$$

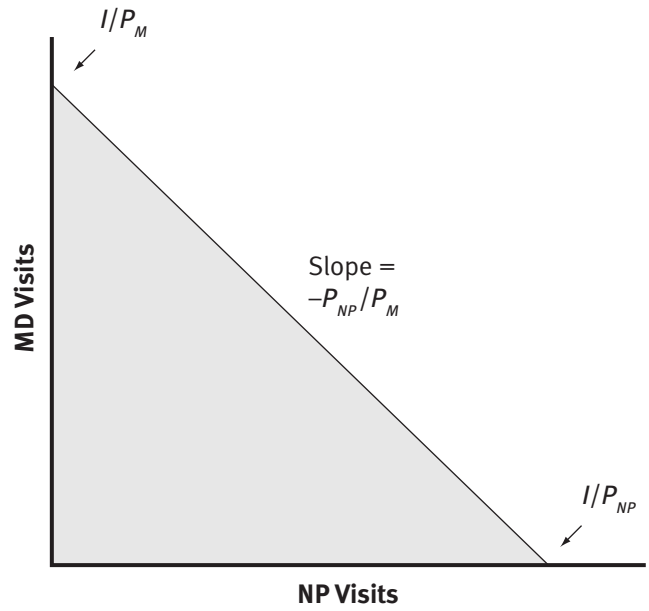
where I is income, Q_M is the quantity of MD services, and Q_{NP} is the quantity of NP services.

Exhibit 2.2 graphs this, using the assumption that all of a person's income during a given period is spent on these two goods. The point at which the budget constraint line intersects each axis shows how many of each good or service the consumer could buy by spending all income on that single service. The first term of Equation 2.4 shows the intercept on the vertical axis, and the term $-P_{NP}/P_M$ is the slope.

The consumer can afford to purchase any combination of these two services that is either on the line or in the shaded area below and to the left of the line, but he cannot afford any combination above and to the right of it. One can easily construct a budget constraint for any level of income because the slope of the line, which is the ratio of the prices of the two goods, does not change. For example, a 50 percent increase in income would shift the budget constraint line outward and to the right by this exact amount.

Also, the slope of the budget constraint is equal to the price ratio between NP and MD visits.[2]

EXHIBIT 2.2
The Budget Constraint



The Consumer Optimum

A rational consumer (defined and discussed in more detail in Chapters 4 and 5) maximizes utility by spending each successive dollar in a way that brings about the most utility. For example, when consumers have spent their last dollar, they will have equalized, across all the goods in their utility function, the ratio of the marginal utilities (MU) with the ratio of the prices (P) of the goods.

If we define P_M as the price of MD visits and P_{NP} as the price of NP visits, then, for a consumer who has maximized her utility,

$$MU_M / P_M = MU_{NP} / P_{NP} \tag{2.6}$$

By cross-multiplying and rearranging the terms, we can write it and think of it another way, where the ratios of the marginal utilities are equal to the price ratios of the two goods:

$$MU_M / MU_{NP} = P_M / P_{NP} \tag{2.7}$$

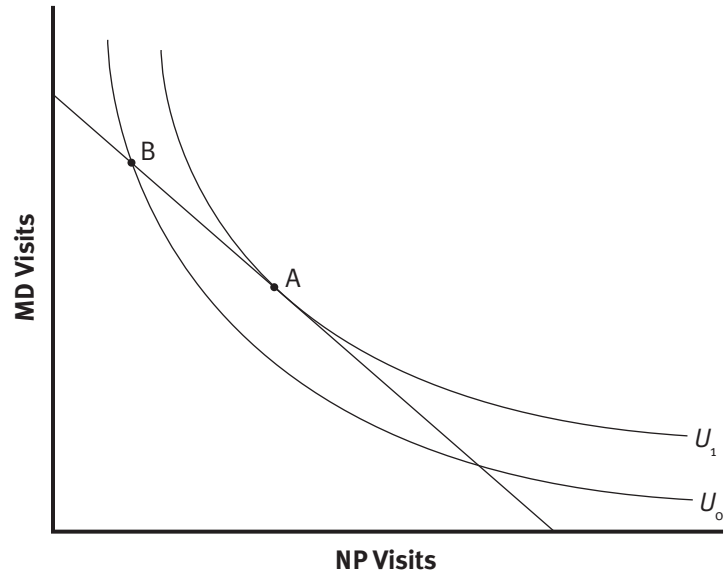
It is easy to see why a consumer must fulfill Equation 2.6 (and therefore Equation 2.7) to maximize utility. Suppose the equality is not met at a particular combination of MD and NP services purchased (which could be point B in Exhibit 2.3). In this case, buying more NP visits and fewer MD visits would benefit the consumer by putting her on a higher indifference curve at point A. The result will be lower marginal utility of NP visits (because of the diminishing marginal utility of the extra NP visits) and higher marginal utility for MD visits. Only when both sides of Equation 2.6 (or 2.7) are equal will the consumer have nothing left to gain from trading one type of visit for another. This trading must be done within the confines of the consumer's budget, however.

Graphically, the consumer chooses the combination of goods that corresponds to the point of tangency between the budget constraint and the highest indifference curve, as illustrated by point A in Exhibit 2.3.

In contrast to point A, at point B the indifference curve and budget constraint do not have the same slope. This fact puts the consumer on an indifference curve that conveys less utility, U_0. By trading MD visits for NP visits, it is possible to move down the budget constraint to point A and increase utility by moving to the higher indifference curve, U_1.

Although much of it may seem obvious, what is remarkable about the theory is that it shows that *everyone* will have the same ratio of marginal utilities between all goods and services. This theory is certainly clear mathematically. If everyone faces the same prices, Equations 2.4 and 2.5 can hold only

EXHIBIT 2.3
The Consumer Optimum



if everyone has the same ratios. But how can that be the case when people exhibit different tastes for alternative goods? Suppose a person prefers NP visits and only wants an occasional MD visit. At the prevailing price ratio, that person will purchase far more NP than MD visits. So at the margin—the last visit—the additional utility of that last NP visit will be relatively low. Moreover, the person uses so few MD services that the last one has a relatively high utility, even though the consumer prefers seeing the NP.

To illustrate, suppose that the price of MD visits is $100 and the price of NP visits is $50. The ratio of the two is two to one. Anyone trying to optimize purchases would make trades until the ratio of the marginal utility for MD visits to the marginal utility of NP visits equaled two to one. This scenario does not imply that each person has the same marginal utility for each type of visit at a particular point on his indifference curve. But the ratio of his marginal utilities at the tangency point to the budget constraint will be two to one.

Consumer theory concludes that, taking into account their own preferences and market prices, people will make the choices that will be most beneficial to them. When people have done as well as they can do, given their resources, they stop trading, presumably to enjoy the goods that they have acquired. These conclusions are strong; much of Chapters 4 and 5 will be devoted to examining and critiquing the assumptions on which they are based.

Demand Curves and Functions

The concept of indifference curves leads naturally to the concept of demand. Here, we develop a demand curve for NP visits. Recall that a consumer has an infinite number of indifference curves, corresponding to the utility received from every possible combination of quantities of the two goods being considered. Exhibit 2.4 shows three indifference curves for NP and MD visits for a particular consumer. Suppose we vary the price of NP visits from P_{NP} to $P_{NP}/2$ to $P_{NP}/4$—that is, we consider what would happen not only at the original price, but also at half and one-fourth that price—but do not change the price of MD visits (P_M) or income. The result is that the budget constraint pivots outward. Under these three alternative sets of prices, we'll assume the consumer chooses to purchase six, eight, and ten NP visits per year, respectively. These points are then plotted as a *demand curve*, labeled D_1 in Exhibit 2.5. (Note that, for simplicity, we show demand and supply curves as straight lines, but there is no reason they cannot have a curvilinear shape.)

A demand curve shows how much of a good is purchased at alternative prices. The curve is drawn under the assumptions that neither the price of other goods nor the person's income changes. Another assumption is that a person's tastes are unaltered. Thus, in deriving the curve, only one thing varies—here, the price of NP services.

Demand curves have a further interpretation: They show the marginal utility a consumer derives from the purchase of the good or service. We

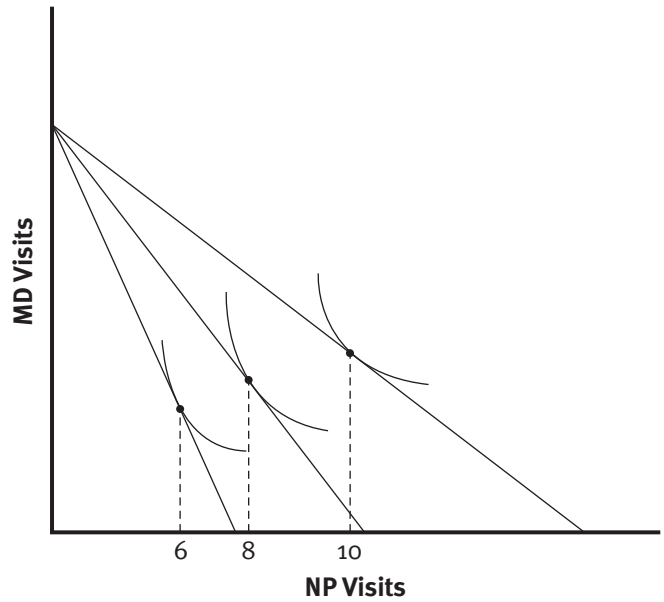
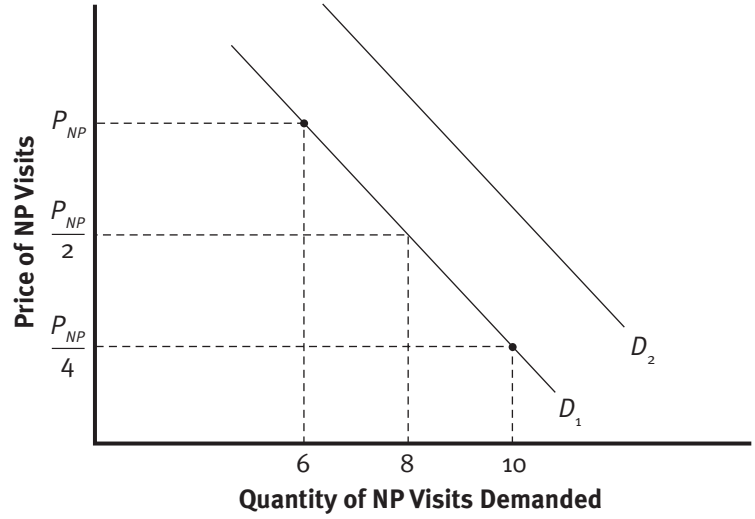


EXHIBIT 2.4 Derivation of Demand Curve, Step 1

EXHIBIT 2.5
Derivation of Demand Curve, Step 2



assume individuals purchase products whose marginal utility exceeds their price. Note that the downward slope of a demand curve follows the decrease in marginal utility as the quantity of a good consumed rises. This topic will be explored further in Chapter 4, where revealed preference and the concept of consumer surplus are discussed.

Although one needs actual data on consumer behavior to draw an accurate demand curve, in general it will slope downward to the right, indicating that people will demand more when the price is lower. In functional form:

$$D = f(P, P_a, I, T) \tag{2.8}$$

where D is demand for a particular good or service, P is its price, P_a is the price of alternatives, I is income, and T is tastes. *Aggregate demand*—how much is demanded by all individuals combined—is simply the sum of the individuals' demands.

Alternative goods, the subscript a, can be categorized two ways: as *complements* or as *substitutes*. Complements are goods that are used in conjunction with the good being studied, and substitutes are goods that are used instead. We can therefore refine Equation 2.8 as follows:

$$D = f(P, P_s, P_c, I, T) \tag{2.9}$$

where P_s is the price of substitutes and P_c is the price of complements.

Most noteworthy about these equations is the unobtrusive role of *T*, tastes. This variable represents much of what it is to be a human being. Psychology and sociology have studied how individual tastes are formed and the ways in which they are manifested. But economic theory takes the taste variable as predetermined and unaffected by the person's environment.

In the health services area, perhaps the major component of *T* relates to health status. If people are sick, they are obviously more likely to use medical care than if they are well. In that sense, they have a "taste" for health. This method would not be a good way to classify your desire for medical services if, say, you were hit by a truck, but there is no other place for it in Equation 2.9. As a result, the demand for health is sometimes expressed as:

$$D = f(P, P_s, P_c, I, HS, T) \tag{2.10}$$

where *HS* is the patient's health status. In Equation 2.10, tastes no longer capture health status, only the non-health-related determinants of demand.

As we saw, a single demand curve can illustrate any relationship between the quantity and price of a particular good. We assume, however, that the other determinants of demand—the prices of alternative goods, income, health status, and tastes—remain unchanged. If they do change, the demand curve must also shift. For example, when a person gets sick, her demand curve is likely to shift outward and to the right—as shown by the demand curve labeled D_2 in Exhibit 2.5—indicating that she will demand more NP visits at all price levels. Nevertheless, the amount demanded is still expected to depend, in part, on the price of NP visits.

The relationships between demand and income and between demand and the price of other goods are more complex. In general, we would expect the demand curve to shift outward and to the right if income rises. This expectation is true of *normal goods*. But there are goods and services with the opposite relationship: When income rises, demand falls, and when income falls, demand rises. These goods and services are known as *inferior goods*—although the term does not imply anything pejorative. A commonly used example of an inferior good is intercity bus travel. As people's income rises, they are likely to use other means of transportation for long-distance travel and use buses less. In the health services area, an example might be visits to an emergency room (ER). People with higher incomes are more likely to have a usual provider of care, and therefore less likely to seek care from an ER. Thus, if income rises we would expect a person's demand curve for ER visits to shift downward and to the left; at any price, the quantity of ER services demanded will be lower.

The relationship between demand for one good and the price of other goods is even more complicated, and we will explore it further when we

discuss elasticities of demand later in the chapter. Two goods are considered substitutes if an increase in the price of one leads to an increase in the demand for the other, and complements if the opposite is true. (A more technical definition, involving cross-price elasticities of demand, is provided later in this chapter.) Most goods and services are substitutes. A classic example is beef and chicken. If the price of beef rises, demand for chicken will increase as people substitute the latter for the former. A classic example of complementary goods is automobiles and tires. If the price of cars rises, demand for cars will decrease, and therefore so will the demand for tires. A health services example of complements is the relationship between inpatient hospital care and outpatient physician services. Although these would seem to be substitutes, there is some evidence to indicate that they are complements: As the price of physician outpatient services rises, the demand for inpatient hospital care falls. (The reason will be explained in Chapter 4.) In summary, if the price of a substitute rises, the demand for the good shifts outward and to the right and the opposite occurs for complements.

Income and Substitution Effects

One of the more challenging concepts in microeconomic theory is income and substitution effects. They are used to illustrate two distinct reasons the quantity of a good or service will tend to rise when the price falls. Suppose the price of MD services falls and the price of NP services remains the same. One reason the quantity of MD services demanded will rise is that relative to the price of NP services, they are now cheaper. People gravitate toward goods that are relatively cheaper than others; this is the substitution effect. The second reason quantity demanded will tend to rise is that, in effect, the reduction in price means the person is no longer spending all of his income. If people use this newfound wealth to purchase more physician services, quantity demanded will rise. This result is the income effect. The total change in the quantity of MD services demanded is the sum of these two effects.

Exhibits 2.6 and 2.7 illustrate the more common case of a normal rather than an inferior good.[3] Exhibit 2.6 shows the expected increase in quantity of MD services demanded, from M_1 to M_2, with a decline in price. Exhibit 2.7 breaks down the increase into income and substitution effects.

Exhibit 2.6 shows two of the consumer's many indifference curves, I_1 and I_2, along with the original, steeper budget constraint and a new, gentler budget constraint corresponding to the reduction in the price of physician services. As the price of physician services falls, the new consumer optimum will shift from point A to point B, corresponding to an increase in the quantity demanded from M_1 to M_2 (illustrated in Exhibit 2.4).

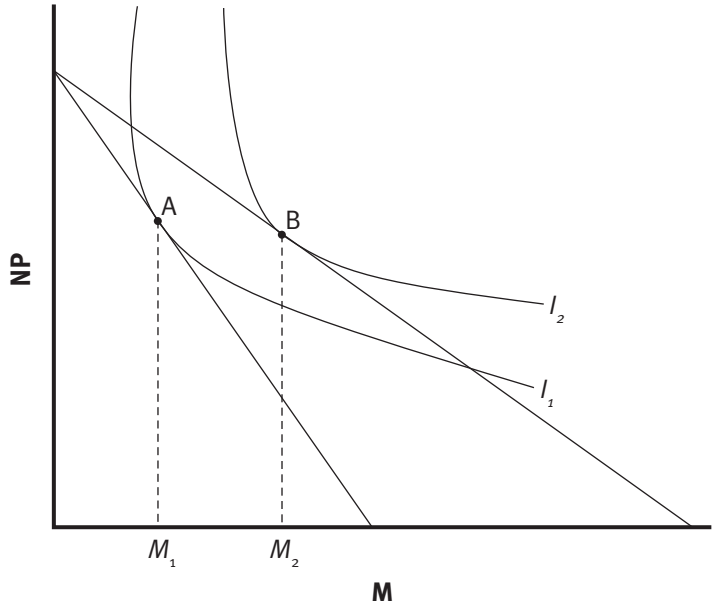


EXHIBIT 2.6
Derivation of Income and Substitution Effects, Step 1

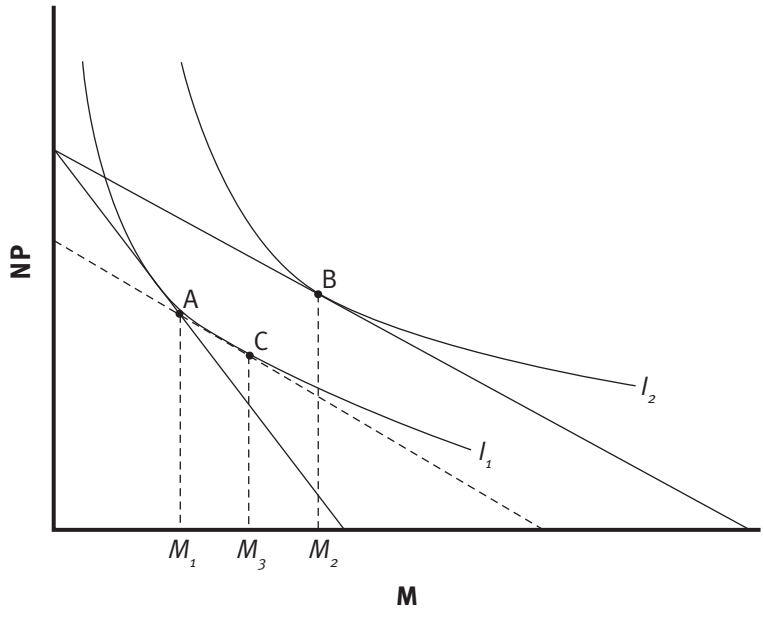


EXHIBIT 2.7
Derivation of Income and Substitution Effects, Step 2

The key to Exhibit 2.7 is the broken line. It runs parallel to the new budget constraint but is closer to the origin. Therefore, the broken line represents the same price ratio as the new budget constraint, because the slope of the budget constraint is equal to the price ratio of the two goods being considered. But note that the broken line is tangent to the original indifference curve, which suggests the consumer is no better off than before the price decrease in MD visits.

Consider consumer optimum point C, where the broken line touches the old indifference curve, and the corresponding quantity purchased, M_3. This point represents how many MD services a person would purchase if that person

- faced the new price ratio and
- was no better off than before prices fell.

The movement from M_1 to M_3 represents the substitution effect—how much the quantity purchased increased solely because of changes in the price ratio, from the original steep line to the new, more gently sloping one. And the movement from M_3 to M_2 shows the income effect—how much the newfound wealth resulting from the lower price of MD services increased the quantity demanded. Their sum, which is the distance from M_1 to M_2, is the total increase in quantity demanded.

Elasticities of Demand

The exact relationship between the quantity of a good purchased and its price is represented by the *price elasticity of demand*. This term is defined as the percentage change in the quantity of a good demanded divided by the percentage change in its price. If the elasticity of demand equals –0.5, this means that when the price of the good changes by, say, 10 percent, the quantity demanded changes by 5 percent, but in the opposite direction. All downward-sloped demand curves have negative signs, so we often drop the sign and refer to its absolute value—here, 0.5.[4]

Health economists have devoted much research to determining demand elasticities for medical services. By convention, goods and services with elasticities exceeding 1.0 are defined as "elastic," those less than 1.0 as "inelastic," and those equaling 1.0 as "unitary elastic." One should not put too much stock in these terms, however. Chapter 4 shows that, although demand elasticities for health services are almost always less than 1.0, they would certainly appear to be price sensitive.

There are three major determinants of the elasticity of a good or service:

- **The extent to which substitutes are available.** If a consumer can easily switch to another good when the price of the original good rises, then the latter will tend to be more elastic.
- **The proportion of the consumer's income spent on the good.** Naturally, one would be more price sensitive about big budget items, such as housing, than tiny ones, such as chewing gum. Thus, goods that comprise a greater share of one's budget tend to have higher elasticities.
- **The time frame in question.** Over time, it's easier for consumers to find substitutes, so long-term elasticities are higher. If the price of gasoline rises, it is hard for consumers to make quick adjustments; gas is price inelastic. Over time, however, persistent high gas prices can lead consumers to lower their demand by buying more fuel-efficient cars, arranging carpools, or using public transportation.

There are two other key elasticities of demand, and in these cases, the sign matters. The *income elasticity of demand* is the percentage change in the quantity of a good demanded divided by the percentage change in a person's income. The income elasticity is calculated in the same way as the price elasticity, substituting income, I, for P. Normal goods have positive income elasticities, and inferior goods have negative ones.

The *cross-price elasticity of demand* is defined as the percentage change in the quantity of a good demanded divided by the percentage change in the price of another good. To return to an earlier example, we might be interested in the cross-price elasticity of demand between inpatient hospital care and outpatient physician services. Substitutes have positive cross-price elasticities, and complements, negative ones.[5]

2.2 Production, Costs, and Supply

Total Product Curves and Isoquants

In production theory, firms seek to maximize profits in the same way consumers attempt to maximize utility. To do so, they purchase inputs and transform them into outputs through the application of some sort of technology. This process is represented using a *production function*.

We will use one of the goods discussed above, NP visits (Q_{NP}), but this time we will examine how a firm produces the good. Assume that several inputs are used to produce these visits through a production process, f. The production function therefore takes the form

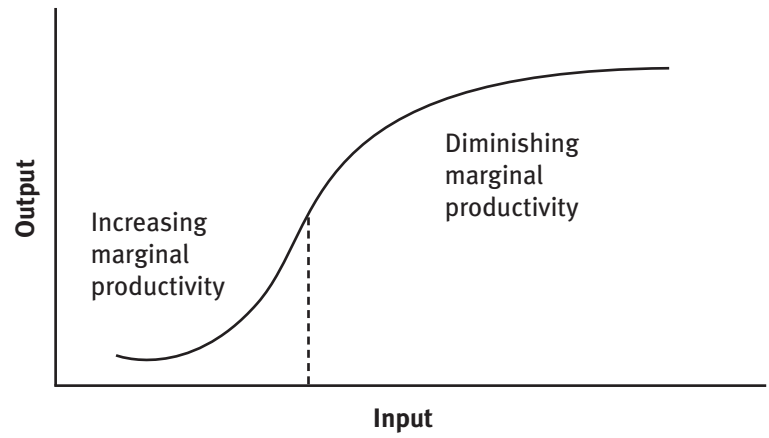
$$Q_{NP} = f(a, b, \ldots, m) \tag{2.11}$$

where *m* inputs, two of which are indicated by the letters *a* and *b*, are used in the production of these visits. The two most important classes of inputs are labor and capital. In the case of NP visits, one would also include supplies (e.g., gloves, dressings, syringes).

The *total product* curve, shown in Exhibit 2.8, shows the relationship between output (the vertical axis) and one particular input (the horizontal axis). Note that when little of the input is used, output increases at an increasing rate, called *increasing marginal productivity*. This rate is a result of the input being underused given the amount of capital available. Loosely, one can think of this as not taking advantage of economies of scale. (The term *economies of scale* has a specific meaning, however, that relates to the long run, and will be defined later.) For example, a single nurse in a big hospital would not be very productive, but as more nurses are added, each will become more productive as each can specialize in particular tasks. In Exhibit 2.8, the rate of additional output eventually decreases as the use of input rises further; this is *diminishing marginal productivity*. It occurs because each new input has less capital (e.g., machines, work space) to use and is therefore less productive than those previously employed.

Returning to Equation 2.11, we will restrict ourselves to two inputs so that these concepts can be represented graphically. Exhibit 2.9 shows curves known as *isoquants*. Quantities of each of two inputs, NPs and examining rooms, are represented on the two axes. The isoquant labeled "Visits = 20" shows the different amounts of inputs required to produce 20 visits per day. The other isoquant indicates the inputs necessary to produce 30 visits per day. The slope of an isoquant at each point is called the *marginal rate of technical substitution* and is equal to the ratios of the marginal productivities of each input at that particular point.[6]

EXHIBIT 2.8
Total Product Curve



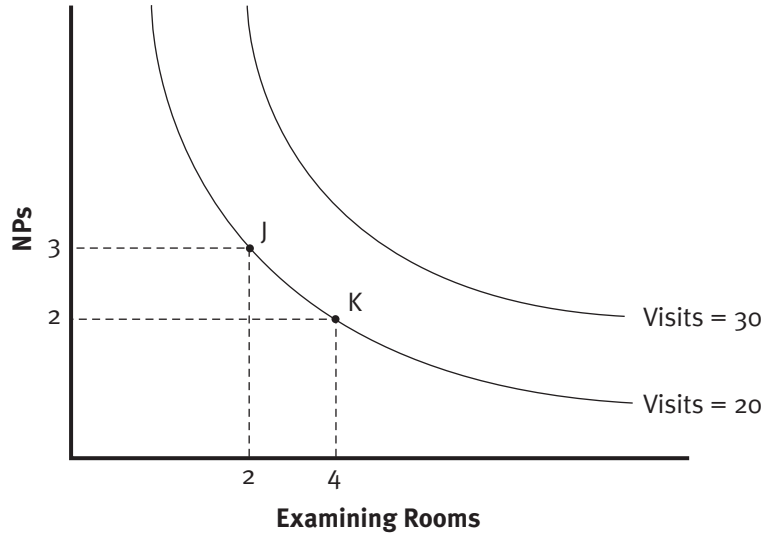


EXHIBIT 2.9
Producer Isoquants

As with indifference curves, isoquants are concave because of diminishing marginal productivity. The *marginal product* is the change in output when a single input is increased by one unit and the other input is held constant. We might expect that a second NP would be able to treat more patients. A third NP would mean even more patients could be treated—but the increase in the number of visits that would result from adding a third NP would likely be smaller than the increase that resulted from adding a second NP. The reason is not that the third NP is necessarily less skilled. Rather, the fixed number of examining rooms creates a physical constraint on the number of patients the office can accommodate. The third NP might help make the use of the examining rooms more efficient, but only so much can be accomplished. If marginal productivity did not diminish, isoquants would be linear.

Why doesn't the practitioner simply get a bigger office, eliminating the constraint on examining rooms? She can, of course. We distinguish between two periods: the *short run* and the *long run*. Over the short run, we assume firms can alter the use of one input (typically labor) but not the other input (usually capital, such as office space). The long run is defined as the period over which a firm can vary all inputs. How long is the long run? Its length depends on what is being produced. It may be short in a simple production process but long for something complicated, such as building new jumbo jets or hospitals. We will return to this concept when we discuss cost curves.

At a given level of technology (represented by our production function, f), only a certain number of visits can be produced. The two isoquants in Exhibit 2.9 indicate that if the state of technology were more advanced, the same number of inputs could produce more outputs. For now, however, we confine ourselves to the lower curve. Points J and K indicate two different ways the office can produce 20 visits per day: with three NPs and two examining rooms, or with two NPs and four examining rooms.

Isocost Lines

Given the state of technology, if a firm produces as much output as possible with a given amount of inputs, production is known to be *technically efficient*. We would expect all firms to strive for this; otherwise, they would not be maximizing profits. It would not necessarily mean, however, that production is *economically efficient*. For economic efficiency, a firm must use the mix of inputs that incurs the least costs. This mix will vary depending on the relative prices of the different inputs. To maximize profits, a firm must be technically and economically efficient.

Input prices are indicated by an *isocost line*, as illustrated in Exhibit 2.10. An isocost line shows how many units of each input the firm can purchase, given their prices and the total amount of money the firm can spend on inputs. The isocost line is analogous to the consumer's budget constraint in several ways: (a) Its point of intersection with each axis shows how many units of that single input can be purchased with the available resources; (b) it is linear (we assume that the market for inputs is competitive and a firm can buy as many as it wishes without affecting the market price); (c) parallel lines that are upward and to the right indicate that the firm has more money to spend on inputs; and (d) the slope of the line is meaningful.

The isocost line can be derived as follows:

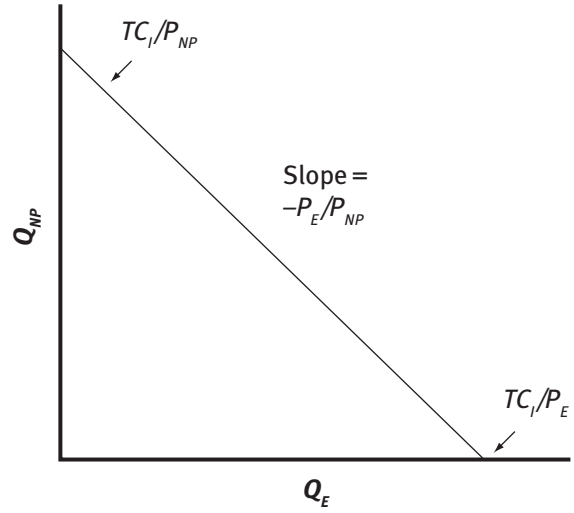
$$TC_I = (P_E \times Q_E) + (P_{NP} \times Q_{NP}) \tag{2.12}$$

Solving for Q_P by rearranging the terms,

$$Q_{NP} = TC_I / P_{NP} - [(P_E / P_{NP}) \times Q_E] \tag{2.13}$$

where TC_I is the total costs of inputs, Q_E is the quantity of examining rooms used, Q_{NP} is the quantity of NPs used, and P_E and P_{NP} are the unit prices of examining rooms and NPs, respectively. The first term of Equation 2.13, (TC_I/P_{NP}), shows the intercept on the vertical axis, and the second term, $-P_E / P_{NP}$, is the slope. The isocost line intersects the horizontal axis at the point TC_I / P_E.

EXHIBIT 2.10
Isocost Line



The Producer Optimum

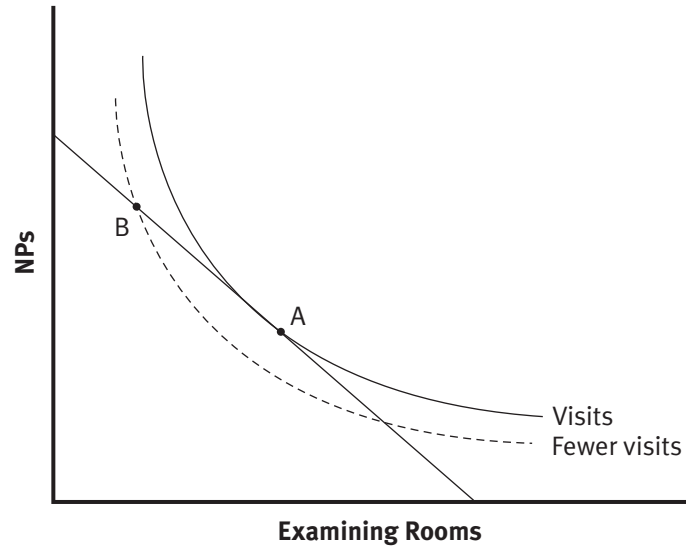
As in the case of consumer theory, the firm achieves its goal—here, profit maximization—at point A in Exhibit 2.11, where the isocost line and isoquant are tangent. At point B, where the lines are not tangent, the production process is economically inefficient: Too many examining rooms and too few NPs are being employed, given the market prices for each of these inputs. Although B is on an isoquant, that isoquant is associated with the production of fewer visits, as indicated by the dashed line.

We saw earlier that the slope of the isocost line is the price ratio between examining rooms and NPs, P_E/P_{NP}. The slope of the isoquant (at a given point) is the ratio of the marginal products of the two inputs. Putting these together, Exhibit 2.11 indicates that a firm will maximize profits when

$$MP_E/MP_{NP} = P_E/P_{NP} \tag{2.14}$$

To see why a firm must fulfill Equation 2.14 to maximize profits, imagine that the equality is not met. The marginal productivity of each input equals 1, but the price of *E* is \$10 and the price of *NP* is \$5. In such a situation, the firm will benefit from hiring more NPs and using fewer examining rooms. Eventually, NPs will become less productive as a result of diminishing marginal productivity. Only when both sides of Equation 2.14 are equal will the firm have no economic reason to change the ratio of inputs it uses. And because all firms face the same input prices, they will all have the same ratio of marginal productivities if they maximize their profits.

EXHIBIT 2.11
The Producer Optimum



Costs of Production in the Short Run

The economic theory of production costs falls neatly out of the theory of production. Recall our assumption that marginal productivity eventually diminishes in the short run, such that increasing the use of one input while capital stock is fixed will increase output, but at a decreasing rate. It follows, then, that the cost of producing each additional unit of output will rise in the short run.

Costs have two components, *fixed costs* and *variable costs*, and their sum is *total costs*. Fixed costs are costs that are invariant with the amount of a good or service produced. One might view the construction of a new hospital wing as a fixed cost. Variable costs, in contrast, rise as more output is produced. NPs would be an example.

Exhibit 2.12 shows the relationship between fixed (*FC*), variable (*VC*), and total (*TC*) costs of production in the short run. Because fixed costs are constant as output increases, they appear as a horizontal line. Variable costs (and therefore total costs) rise with output. At the beginning, when the input is underused, they rise at a decreasing rate, but eventually they rise at an increasing rate. This rate is consistent with the notion of diminishing marginal productivity.

We will consider two more types of curves, representing *average costs* and *marginal costs*. Average costs (*AC*) are simply total costs divided by the number of units of output produced. As shown in Exhibit 2.13, average

fixed costs (*AFC*) will always fall as output increases because fixed costs are constant. Average variable costs (*AVC*) and average total costs (*ATC*) fall initially, when more use of the input is more economically efficient, but eventually they begin to rise as diminishing marginal productivity sets in.

The other curve in Exhibit 2.13, *marginal costs* (*MC*), is key in microeconomic theory. Marginal costs are the change in total costs when one more unit of output is produced. Note that marginal costs intersect the minimum points of the average variable costs and average total cost curves. If marginal costs are lower than the average, that last unit produced will pull the average down. If marginal costs are higher than the average, the additional unit produced will pull the average up. But when the marginal and average costs are equal, the last unit produced will not change the average, and average costs will be flat at that point.

A number of factors affect the position of the cost curves. If inputs are more expensive, cost curves will shift upward, reflecting higher total and marginal costs at any given level of output. If a cost-saving technology is developed, the curves will be lower, because it will cost less to produce a given output. Finally, if the quality of output increases, costs will also be affected. In general, it costs more to produce higher quality.

Long-Run Costs and Economies of Scale

Although we will not touch on it again in this chapter, an understanding of the concept of long-run costs is important. Recall that the long run is

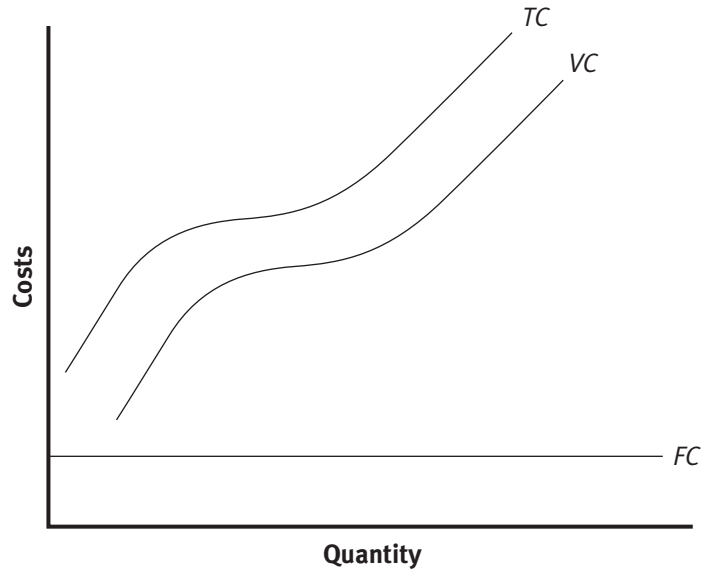
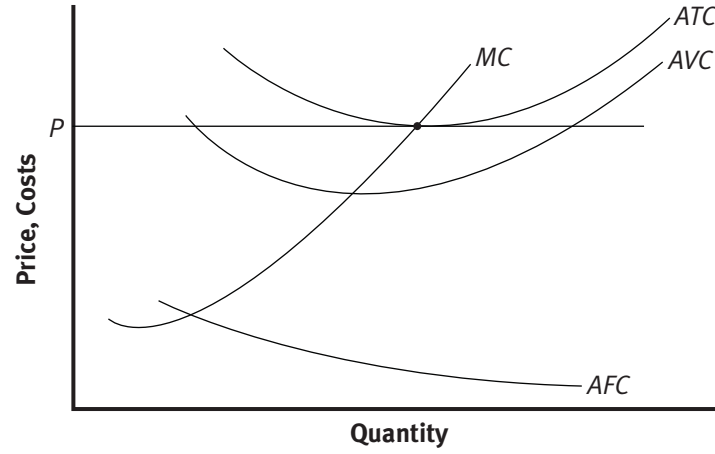


EXHIBIT 2.12
Total, Variable, and Fixed Costs

EXHIBIT 2.13
Average and Marginal Costs in the Short Run



the period over which a firm can vary all of its inputs. The curve shown in Exhibit 2.14 represents *long-run average costs* (*LRAC*), the average costs of producing any given level of output over the long run. *Economies of scale* are situations in which output rises at a greater rate than does the increase in the cost of inputs. For example, a firm increases its use of all inputs by 10 percent, which results in a 15 percent increase in output. The opposite situation would represent a *diseconomy of scale*: in this example, output rising by, say, 5 percent. Situations in which the costs of inputs and outputs rise by the same amount are called *constant returns to scale*. All points to the left of Q^* show areas where a firm experiences economies of scale, and to the right, diseconomies.

Firms can experience economies of scale for several reasons, including specializing, purchasing inputs in greater numbers, and using advertising or labor more efficiently. But why would a firm eventually reach a diseconomy of scale, given that in the long run, more capital can always be purchased? The major cause is managerial issues. Eventually, a firm gets too big and unwieldy to function optimally. A 3,000-bed hospital would probably be far more difficult to manage than one half its size.

The relationship between the short-run average cost curve shown in Exhibit 2.13 and long-run average costs in Exhibit 2.14 is more complicated. Recall that by definition, in the short run the firm is constrained by a fixed stock of capital. As a result, its costs can never be lower than the firm would face in the long run, when it can vary the use of all inputs.

The long-run average cost curve, shown again in Exhibit 2.15, is the envelope of all possible short-run average cost curves. In theory, there is one

short-run average cost curve that corresponds to the cheapest way to produce a given level of output at each possible level of capital. In the exhibit, $SRAC_1$ shows short-run costs when there is a small amount of capital. Average costs would be minimized if the firm desired to produce a quantity of Q_1 but very high if the firm wanted to produce more. Similarly, we see when examining $SRAC_3$ that this much larger plant size would be optimal for producing Q_3 of the good. However, there very well might not be sufficient demand for

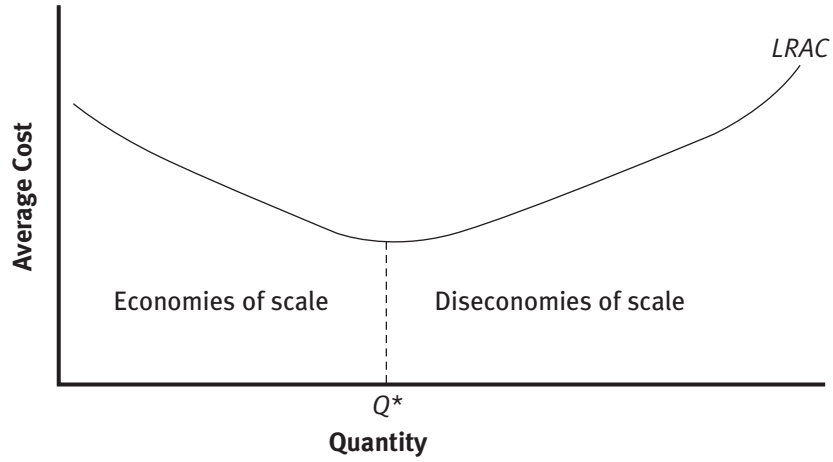


EXHIBIT 2.14
Long-Run Average Costs

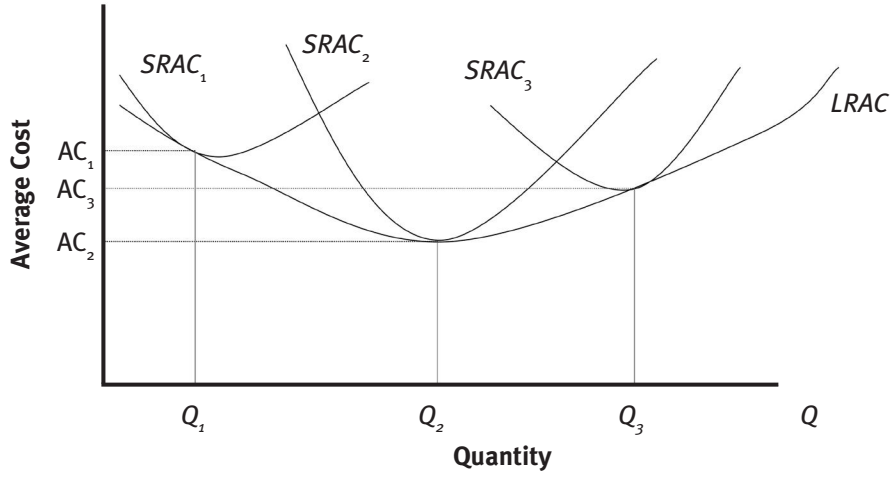


EXHIBIT 2.15
Relationship Between Short-Run and Long-Run Average Costs

the good to sell such a large quantity at a profitable price. For lower levels of output, average costs will be high. As drawn, a middle-sized plant lowers costs the most, as shown by $SRAC_2$.

Applying this to health, imagine a hospital with ten beds; it probably could not produce care as cheaply, on average, as a bigger hospital. Its lowest costs are at AC_1, and by expanding it could achieve economies of scale. The opposite is true at the highest level of capital; the facility is just too big. Suppose a hospital has 3,000 beds. At this point, with an average cost of AC_3, it is experiencing diseconomies of scale. Average costs are minimized at AC_2, when the hospital is at the middle size.

The concept of long-run average costs is also more complicated because one considers it in the planning stages of a production process. A firm has to decide what size capital outlay to make. Once it makes this outlay, it is "stuck" in the short run. If production is planned poorly, costs will be higher than necessary until the firm can change its capital stock again.

How a Competitive Firm Chooses a Profit-Maximizing Level of Output

We return now to the short run. After choosing the most economically efficient mix of inputs, firms must decide how much to produce and the price for which they will sell their output. In a competitive market, in which the products of alternative firms are indistinguishable, there really is no choice regarding price: A firm will lose all of its market if it charges more than the going price, and it will not maximize profits if it charges less. (We discuss how this market price is determined in Section 2.3.)

A competitive (as opposed to monopolistic) firm chooses the quantity to produce by equating the marginal cost (*MC*) of production—the cost of producing the last good—to the market price of the good. If it produces less, it is not maximizing profits, because it would make more money by producing more units. If it produces more, it loses money on the last units produced. Point A in Exhibit 2.16 shows the optimal production amount, with a corresponding quantity Q produced at price P. The firm faces a fixed, or horizontal, market price for selling the good, but the marginal cost curve slopes upward, reflecting the likelihood that it will cost more and more to produce successive units of output.

Derivation of the Supply Curve

The *supply curve* shows how much a profit-maximizing firm will produce at different prices. In general, we expect it to slope upward, indicating that firms will produce more if they can receive more from selling their product.

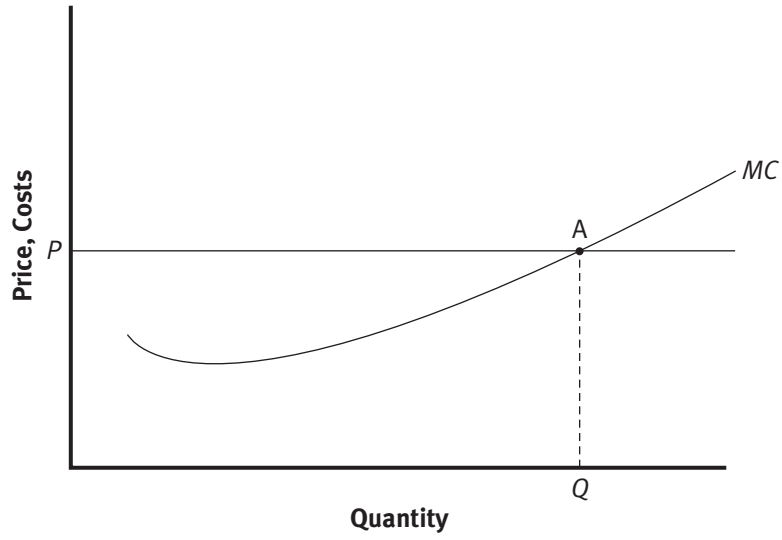


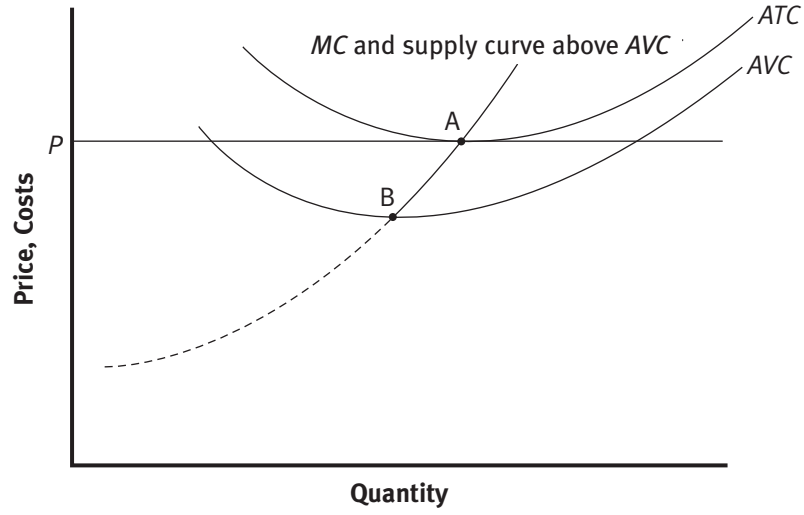
EXHIBIT 2.16 How Competitive Firms Choose Output Levels

An individual firm's short-run supply curve is its marginal cost curve at all points above average variable costs (*AVC*), as shown in Exhibit 2.17. All points above and to the right of point B on this curve show how much a profit-maximizing firm would be willing to produce at different price levels. We could clarify this if we were to plot several horizontal lines that intersect *MC*, indicating alternative market prices. Each of the intersections represents a profit-maximizing level of output. *Aggregate supply*—the total quantity supplied in a market—is simply the sum of each firm's individual supply.

Imagine a market price that is below and to the left of *AVC*, on the broken portion of the curve. A firm would never produce there because the costs of its labor and supply inputs would exceed the price—it would lose money. A firm *would* be willing to produce at all points above the average total cost curve (*ATC*)—above and to the right of point A—because with prices that high, it would make a profit on each item it produced. As a result, the section of the *MC* curve above *ATC* is part of the firm's supply curve.

Why would a firm be willing to produce at points on the *MC* curve that are *between AVC* and *ATC* (between points A and B)? The firm would not lose money—the price is high enough to cover labor and supplies, the main components of *AVC*. Thus, the firm would be willing to stay in business, at least in the short run, because proceeds cover expenses. In the long run, however, the price is not sufficient to pay off the firm's fixed costs.

EXHIBIT 2.17
Derivation of a Firm's Supply Curve



To fully explain these concepts, we must define *economic costs* and *economic profits*. One normally thinks of costs as those expenses associated with production, but in economics, the standard definition includes *opportunity costs* as well. Opportunity costs are usually defined as the value of the next best opportunity. Thus, the value of investing one's resources in something else that would provide a return—for example, the interest rate associated with investments—would be a component of costs. As a result, we include this forgone opportunity as a cost of doing business. Thus, economic costs exceed what an accountant might define as the cost of doing business.

In a competitive market, a firm's profits are *zero*. Certainly, firms wish to make a higher profit, but if profits are to be made, other firms will enter the market, which will lower price and bring profits back to zero. So how can a firm survive on zero economic profits? It goes back to the definition: A zero profit to an economist is a positive profit to an accountant, because economists consider a normal rate of return part of costs. If the normal rate of return is 5 percent per year, then accounting profits of 5 percent would be equivalent to zero economic profits.

The prices of inputs and technology can cause a supply curve to shift.[7] We can therefore write the firm's supply schedule as:

$$S = f(P, P_i, Tech) \tag{2.15}$$

where S is the amount of the good supplied, P is its market price, P_i is the price of inputs, and *Tech* is the level of technology.

Suppose a technological breakthrough reduces the cost of production. Then, at any price, the firm could profitably produce more. Alternatively, at any quantity, the costs of production would be lower. The supply curve would shift outward and to the right. If the cost of inputs declines, the supply curve would shift in the same direction, for the same reason. However, most analysts in the health services area find technology to be, in general, cost increasing rather than cost decreasing (Newhouse 1993b). If this is true, these technologies are probably designed to improve people's health status or comfort, not to reduce costs.

Just as there are demand elasticities, there is a *price elasticity of supply*. The price elasticity of supply is calculated in the same manner as the price elasticity of demand, replacing the quantity consumers demand with the quantity firms supply. In contrast to demand elasticities, supply elasticities tend to be positive, reflecting the positive slope of the supply curve.

In summary, production theory predicts that firms will seek to use their inputs as efficiently as possible and make their output choices in a way that maximizes their profits. Doing so also serves social purposes in that firms are (1) not wasting inputs and (2) only producing those goods and services that consumers demand. The next section describes the interaction of the many consumers and products that make up the economy as a whole.

2.3 Equilibrium in a Competitive Market

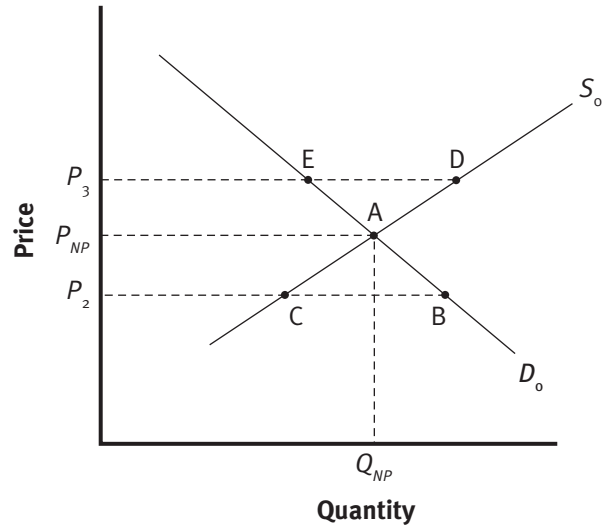
Short-Run Equilibrium

The combination of demand and supply determines the price and output levels in a market, but we must also consider aggregate demand and supply—that is, all consumers and firms combined for a particular good or service.

In Exhibit 2.18, the price, P_{NP}, and quantity, Q_{NP}, for a particular market (here, NP services) are in equilibrium, shown as point A, where the demand and supply curves intersect. If the price were lower, say P_2, the quantity demanded at point B would be greater than the quantity supplied at point C. This situation is defined as a *shortage*. In contrast, if the price were higher than the equilibrium level, at P_3, then supply at point D would exceed demand at point E, a situation defined as a *surplus*. Only at the equilibrium price P_{NP} is there no shortage or surplus.

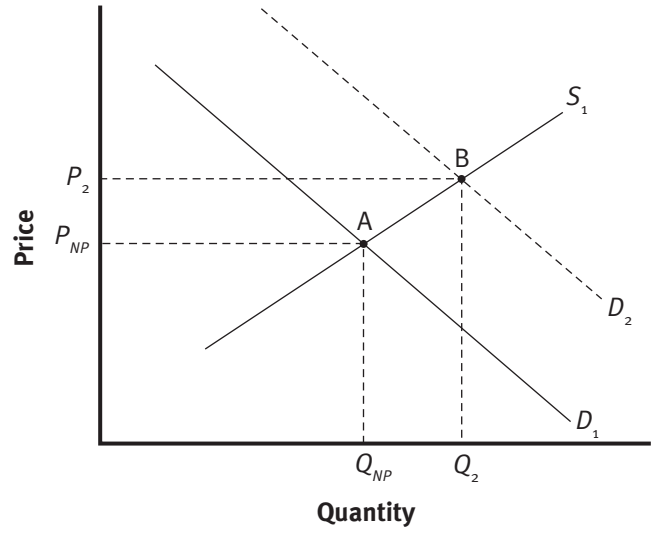
In the short run, a variety of factors can disturb this equilibrium. One example is an increase in demand, shown in Exhibit 2.19, caused perhaps by an increase in income. If we assume NP services are a normal good (i.e., more is demanded when income is higher), this would cause an upward shift to the right in the demand curve, as shown by the broken demand curve, D_2. The

EXHIBIT 2.18 Short-Run Equilibrium in a Competitive Market



new equilibrium would be at point B, which corresponds to both a higher price and a higher quantity, where the new (broken) demand curve intersects the old (solid) supply curve. Supply can also shift in the short run. For example, this would occur if the cost of an important input rose.

EXHIBIT 2.19 Shifts in Demand and Supply in the Short Run



Long-Run Equilibrium

In the long run, firms can adjust all inputs, including capital. Because of this, new firms can go into and out of business. Perhaps the easiest way to understand this concept is to think of three sequential periods. Period 1 is the baseline, as shown by the equilibrium in Exhibit 2.18. Period 2 is the short run, illustrated in Exhibit 2.19. Period 3 is the long run.

Suppose demand for NPs increases as shown by the broken demand curve in Exhibit 2.19. This demand causes the price to increase. In the long run, shown in Exhibit 2.20, there is time for a supply response. In this case, we would expect more NPs to enter the market because the profession would become more lucrative. This scenario might seem far-fetched, because it takes a long time to train more of these professionals, but it is actually feasible. There also may be many individuals with such training who, for whatever reason, are currently not in the job market or are working in another profession. The higher prices for their services could stimulate them to reenter the market.

In Exhibit 2.20, the new equilibrium, where the broken demand and supply curves intersect, point C, is at the original price, but at a much higher quantity than before, Q_3. The logic is straightforward. The original price in Exhibit 2.18 was disturbed by an increase in demand, shown in Exhibit 2.19. Because the quantity demanded exceeded the amount supplied, the price rose to P_2. But this increase in price stimulated an increase in supply, as shown in Exhibit 2.20. Because NPs were less scarce, firms were able to pay

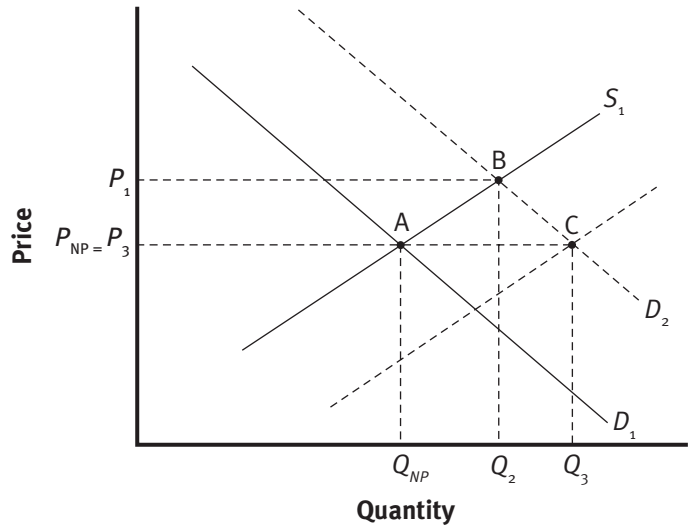


EXHIBIT 2.20 Shifts in Demand and Supply in the Long Run

a lower price. In reality, the price in Exhibit 2.20 might be a little higher or lower than that in Exhibit 2.18. This price would depend largely on whether there were increasing or decreasing economies of scale in the production of NP services.

2.4 Equilibrium for a Monopolist

Strictly speaking, a *monopoly* exists when one firm supplies all of the goods or services in a market. (A *monopsony*, in contrast, occurs when there is a single purchaser rather than producer.) There are milder forms of monopoly as well: a *duopoly* exists when two firms supply an entire market, and an *oligopoly* when just a few supply the market. We will focus on monopoly.

Sometimes we speak of a firm having "monopoly power" or "market power." This term does not mean there is only one firm in the market. Rather, it means the firm has some ability to raise its price without losing its entire market. That may seem odd, if you recall that a competitive firm faces a horizontal demand curve for its product—if it charged more, it would no longer have any customers. Graphically, monopoly power means the demand curve a firm faces has a somewhat downward slope.[8] Most firms do have some monopoly power, and retaining this power is one of the main purposes of advertising. Consider Cheerios, for example. There are many substitutes, but if the price of Cheerios goes up, some people will still buy them, albeit fewer than before.

Monopolists Charging a Single Price

There are two classes of monopolists: those that charge everyone the same price and those that can charge different prices to different customers. An example of the former is difficult to find in the health sector, as most monopolists—for example, prescription drug companies—are able to charge different prices to different customers, as they often make (confidential) deals with health insurance and managed care companies. Nevertheless, the logic is important to understand.

Consider the concept of *marginal revenue*, which is the amount of money brought in by the sale of the last unit of a good. In a competitive market, a marginal revenue curve (with quantity on the horizontal axis and marginal revenue on the vertical axis) is flat and is equal to the price. This curve, in turn, is equal to the demand curve facing the firm. No matter how many units a particular firm sells, it receives the same market price or marginal revenue for each one. The reason is that each firm is assumed to be such a small part of the market that its sales have no effect on overall market price. An example might be a soybean farmer.

This case does not apply to a monopolist, however. Because a monopolist *is* the entire market, the demand curve it faces is not horizontal, but downward sloping. If it charges more, demand will decrease.

Recall that demand represents how much people will pay for a product at a particular price. In a competitive market, price, demand, and marginal revenue are all equal. If the market price of a good is $20, that is the demand the firm faces, or how much buyers are willing to pay. If the firm sells an extra unit, it receives an extra $20 in marginal revenue. This scenario is not the case with a monopolist. True, if it sells another unit of a good, it receives more money. However, as we are considering a monopolist that charges a single price, it must lower the price for all units sold to sell an extra unit. As a result, marginal revenue is lower than demand because the extra money it receives from the last unit is reduced by the lower price on all previously sold units.

As a result, the marginal revenue curve will always be lower than the demand curve, as illustrated in Exhibit 2.21. Consider the marginal revenue for the eleventh unit sold. If 10 units are sold, suppose the monopolist can charge $21 each, but if it wants to sell 11, it can only charge $20.Therefore, the total revenue for 10 units is $210, and for 11 units, $220. The marginal revenue is just $10—well below the demand curve, which at a quantity of 11 units is $20. Therefore, a monopolist's marginal revenue curve must be lower than its demand curve.

Exhibit 2.22 shows how much a monopolist charges and produces to maximize its profits. The rule of thumb is that it produces where its marginal costs and marginal revenue are equal, at point A, corresponding to a quantity of Q_M. It then chooses the highest price it can charge for this quantity, which

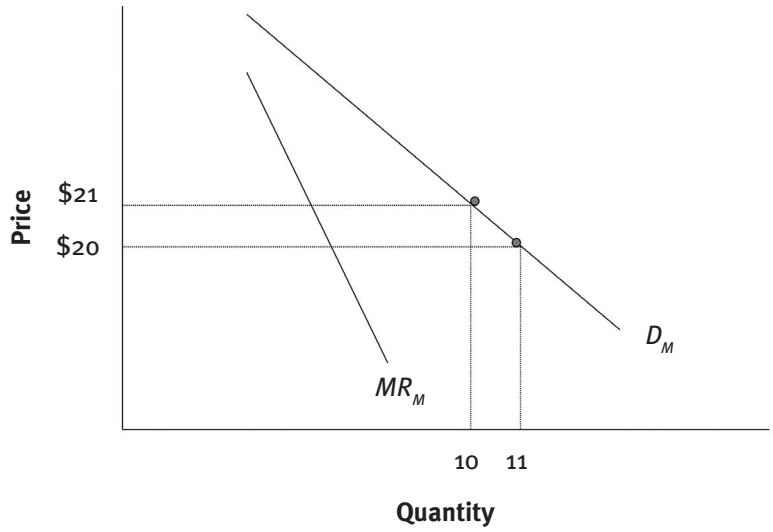


EXHIBIT 2.21 Calculation of Marginal Revenue for a Monopolist

is found by locating the point on the demand curve corresponding to this quantity, P_M. What a competitive firm would produce and charge is superimposed on the graph, at Q_C and P_C (the intersection of demand and supply). The monopolist charges more and produces less than the competitive firm. To maximize its profits, a monopolist can charge more than a competitive firm. But by charging more, there is insufficient demand to sell as many units as in a competitive market.

A monopolistic market is disadvantageous to society in two respects: Customers pay more, and less is produced for society. As a result, *antitrust* laws are aimed at preventing the formation and continuation of some monopolies. There are, however, certain situations in which the market will not support more than one firm, necessitating a monopoly. Transportation systems, such as bus, subway, or freeway systems, are possible examples, as are electrical utility systems. In such a case, which we refer to as a *natural monopoly*, government may either take over the market (subways) or regulate it (utilities) to ensure that the private-firm monopolist doesn't charge more than is necessary.

Price-Discriminating Monopolist

A price-discriminating monopolist can charge different prices to different customers. Although one can draw a graph illustrating this, it is easier to understand through a description.

EXHIBIT 2.22 Equilibrium for a Monopolist Charging a Single Price