eleventh edition

The Economic Way of Thinking

PAUL HEYNE · PETER BOETTKE · DAVID PRYCHITKO

七彩英语
QCENGLISH.COM

CHAPTER 1

Microeconomics, A Way of Thinking about Business

In economics in particular, education seems to be largely a matter of unlearning and "disteaching" rather than constructive action. A onc- famous American humorist observed that "it's not ignorance that does so much damage; it's knowin' so darn much that ain't so." . . .It seems that the hardest things to learn and to teach are things that everyone already knows.

Frank H. Knight

Frank Knight was a wise professor. Through long years of teaching he realized that students, even those in advanced business programs, beginning a study of economics, no matter the level, face a difficult task. They must learn many things in a rigorous manner that, on reflection and with experience, amount to common sense. To do that, however, they must set aside —or "unlearn"—many pre-conceived notions of the economy and of the course itself. The problem of "unlearning" can be especially acute for MBA students who are returning to a university after years of experience in industry. People in business rightfully focus their attention on the immediate demands of their jobs and evaluate their firms' successes and failures with reference to production schedules and accounting statements, a perspective that stands in stark contrast to the perspective developed in an economics class.

As all good teachers must do, we intend to challenge you in this course to rethink your views on the economy and the way firms operate. We will ask you to develop new methods of analysis, maintaining all the while that there is, indeed, an "economic way of thinking" that deserves mastering. We will also ask you to reconsider, in light of the new methods of thinking, old policy issues, both inside and outside the firm, about which you may have fixed views. These tasks will not always be easy for you, but we are convinced that the rewards from the study ahead are substantial. The greatest reward may be that this course of study will help you to better understand the way the business world works and how businesses might be made more efficient and profitable. Much of what this course is about is, oddly enough, crystallized in a story of what happened in a prisoner-of-war camp.

The Emergence of a Market

Economic systems spring from people's drive to improve their welfare. R.A. Radford, an American soldier who was captured and imprisoned during the Second World War, left a vivid account of the primitive market for goods and services that grew up in his prisoner-of-war camp.[1] A **market** is the process by which buyers and sellers determine what they

[1] R.A. Radford, "The Economic Organization of a POW Camp," *Economica* (November 1945), pp. 180-201.

are willing to buy and sell and on what terms. That is, it is the process by which buyers and sellers decide the prices and quantities of goods that are to be bought and sold. Because the inmates had few opportunities to produce the things they wanted, they turned to a system of exchanges based on the cigarettes, toiletries, chocolate, and other rations distributed to them periodically by the Red Cross.

The Red Cross distributed the supplies equally among the prisoners, but "very soon after capture . . .[the prisoners] realized that it was rather undesirable and unnecessary, in view of the limited size and the quality of supplies, to give away or to accept gifts of cigarettes or food. Goodwill developed into trading as a more equitable means of maximizing individual satisfaction."[2] As the weeks went by, trade expanded and the prices of goods stabilized. A soldier who hoped to receive a high price for his soap found he had to compete with others who also wanted to trade soap. Soon shops emerged, and middlemen began to take advantage of discrepancies in the prices offered in different bungalows.

A priest, for example, found that he could exchange a pack of cigarettes for a pound of cheese in one bungalow, trade the cheese for a pack and a half of cigarettes in a second bungalow, and return home with more cigarettes than he had begun with. Although he was acting in his own self-interest, he had provided the people in the second bungalow with something they wanted—more cheese than they would otherwise have had. In fact, prices for cheese and cigarettes differed partly because prisoners had different desires and partly because they could not all interact freely. To exploit the discrepancy in prices, the priest moved the camp's store of cheese from the first bungalow, where it was worth less, to the second bungalow, where it was worth more. Everyone involved in the trade benefited from the priest's enterprise.

A few entrepreneurs in the camp hoarded cigarettes and used them to buy up the troops' rations shortly after issue—and then sold the rations just before the next issue, at higher prices. An **entrepreneur** is an enterprising person who discovers potentially profitable opportunities and organizes, directs, and manages productive ventures. Although these entrepreneurs were pursuing their own private interest, like the priest, they were providing a service to the other prisoners. They bought the rations when people wanted to get rid of them and sold them when people were running short. The difference between the low price at which they bought and the high price at which they sold gave them the incentive they needed to make the trades, hold on to the rations, and assume the risk that the price of rations might not rise.

Soon the troops began to use cigarettes as money, quoting prices in packs or fractions of packs. (Only the less desirable brands of cigarette were used this way; the better brands were smoked.) Because cigarettes were generally acceptable, the soldier who wanted soap no longer had to search out those who might want his jam; he could buy the soap with cigarettes. Even nonsmokers began to accept cigarettes in trade.

This makeshift monetary system adjusted itself to allow for changes in the money supply. On the day the Red Cross distributed new supplies of cigarettes, prices rose, reflecting the influx of new money. After nights spent listening to nearby bombing, when

---

[2] Ibid., pg. 190.

the nervous prisoners had smoked up their holdings of cigarettes, prices fell. Radford saw a form of social order emerging in these spontaneous, voluntary, and completely undirected efforts. Even in this unlikely environment, the human tendency toward mutually advantageous interaction had asserted itself.

Today, markets for numerous new and used products spring up spontaneously in much the same way. At the end of each semester, college students can be found trading books among themselves, or standing in line at the bookstore to resell books they bought at the beginning of the semester. Garage sales are now common in practically all communities. Indeed, like the priest in the POW camp, many people go to garage sales to buy what they believe they can resell—at a higher price, of course. "Dollar stores" have sprung up all over the country for one purpose, to buy the surplus merchandise from manufacturers and to unload it at greatly reduced prices to willing customers. There are even firms that make a market in getting refunds for other firms on late overnight deliveries. Many firms don't think it is worth their time to seek refunds for late deliveries, mainly because there aren't many late deliveries (because the overnight delivery firms have an economic incentive to hold the late deliveries in check). However, there are obviously economies to be had from other firms collecting the delivery notices from several firms and sorting the late ones out with the refunds shared by all concerned.

Today, we stand witness to what is an explosion of a totally new economy on the Internet that many of the students reading this book will, like the priest in the POW camp, help develop. More than two hundred years ago, Adam Smith outlined a society that resembled these POW camp markets in his classic *Wealth of Nations* (see the "Perspective" on Smith page after next). Smith, considered the first economist, asked why markets arise and how they contribute to the social welfare. In answering that question, he defined the economic problem.

The Economic Problem

Our world is not nearly as restrictive as Radford's prison, but it is no Garden of Eden, either. Most of us are constantly occupied in securing the food, clothing, and shelter we need to exist, to say nothing of those things we would only like to have—a tape deck, a night on the town. Indeed, if we think seriously about the world around us, we can make two general observations.

First, the world is more or less fixed in size and limited in its resources. **Resources** are things used in the production of goods and services. There are only so many acres of land, gallons of water, trees, rivers, wind currents, oil and mineral deposits, trained workers, and machines that can be used in any one period to produce the things we need and want. We can plant more trees, find more oil, and increase our stock of human talent, but there are limits on what we can accomplish with the resources at our disposal.

Economists have traditionally grouped resources into four broad categories: land, labor, capital (also called investment goods), and technology.[3] To this list some economists would add a fifth category, entrepreneurial talent. The entrepreneur is critical to the success of any economy, especially if it relies heavily on markets. Because entrepreneurs discover more effective and profitable ways of organizing resources to produce the goods and services people want, they are often considered a resource in themselves.

Our second general observation is that in contrast to the world's physical limitations, human wants abound. You yourself would probably like to have books, notebooks, pens and a calculator, perhaps even a computer with a gigabyte worth of RAM and an 80 gigabyte hard-disk drive. A stereo system, a car, more clothes, a plane ticket home, a seat at a big concert or ballgame—you could probably go on for a long time, especially when you realize how many basics, like three good meals a day, you normally take for granted.

In fact, most people want far more than they can ever have. One of the unavoidable conditions of life is the fundamental condition of scarcity. **Scarcity** is the fact that we cannot all have everything we want all the time. Put simply, there isn't enough of everything to go around. Consequently, society must face several unavoidable questions:

1. What will be produced? More guns or more butter? More schools or more prisons? More cars or more art, more textbooks or more "Saturday night specials"?
2. How will those things be produced, considering the resources at our disposal? Shall we use a great deal of labor and little mechanical power, or vice versa? And how can a firm "optimize" the use of various resources, given their different prices?
3. Who will be paid what and who will receive the goods and services produced? Shall we distribute them equally? If not, then on what other basis shall we distribute them?
4. Perhaps most important, how shall we answer all these questions? Shall we allow for individual freedom of choice, or shall we make all these decisions collectively?

These questions have no easy answers. Most of us spend our lives attempting to come to grips with them on an individual level. What should I do with my time today—study or walk through the woods? How should I study—in the library or at home with the stereo on? Who is going to benefit from my efforts—me or my mother, who wants

---

[3] Land includes the surface area of the world and everything in nature—minerals, chemicals, plants—that is useful in the production process. Labor includes any way in which human energy, physical or mental, can be usefully expended. Capital (investment goods) includes any output of a production process that is designed to be used later in other production processes. Plants and equipment—things produced to produce other things—are examples of these manufactured means of production. Technology is the knowledge of how resources can be combined in productive ways.

me to succeed? Am I going to live by principle or by habit? Take each day as it comes or plan ahead? In a broader sense, these questions are fundamental not just to the individual but to all the social sciences, economics in particular. Scarcity is the root of economics. **Economics** is the study of how people cope with scarcity—with the pressing problem of how to allocate their limited resources among their competing wants in order to satisfy as many of those wants as possible. More to the point, it is a way of *thinking* about how people, individually and collectively in various organizations (including firms), cope with scarcity.

The problem of allocating resources among competing wants is not as simple as it may first appear. You may think that economics is an examination of how one person or a small group of people makes fundamental social choices on resource use. That is not the case. The problem is that we have information about our wants and the resources at our disposal that may be known to no one else. This is a point the late Leonard Reed made decades ago in a short article in terms of what it takes to produce a product as simple as a pencil (see the reading "I, A Pencil" at the end of the chapter), and it also is a point that F. A. Hayek stressed throughout all of his writings that, ultimately, gained him a Nobel Prize in economics (see the reading "The Use of Knowledge in Society" in your course packet). For example, you may know you want a calculator because your statistics class requires you to have one, and even your friends (much less the people at Hewlett-Packard or Casio) do not yet know your purchase plans. You may also be the only person who knows how much labor you have, which is determined by exactly how long and intensely you are willing to work at various tasks. At the same time, you may know little about the wants and resources that other people around the country and world may have. Before resources can be effectively allocated, the information we hold about our individual wants and resources must somehow be communicated to others. This means economics must be concerned with systems of communications. Indeed, the field is extensively concerned with how information about wants and resources is transmitted or shared through, for example, prices in the market process and votes in the political process. Indeed, the "information problem" is often acute within firms, given that the CEO often knows little about how to do the jobs at the bottom of the corporate "pyramid." The information problem is one important reason that firms must rely extensively on incentives to get their workers (and managers) to pursue firm goals.

Markets like the one in the POW camp and even the firms that operate within markets emerge in direct response to scarcity. Because people want more than is immediately available, they produce some good and services for trade. By exchanging things they like less for things they like more, they reallocate their resources and enhance their welfare as individuals. As we will see, people organize firms, which often substitute command-and-control structures for the competitive negotiations and exchanges of markets, because the firms are more cost-effective than markets. Firms can be expected to expand only as long as they remain more cost-effective than competitive market trades.

The Scope of Economics

MBA students often associate economics with a rather narrow portion of the human experience: the pursuit of wealth; money and taxes; commercial and industrial life. Critics often suggest that economists are oblivious to the aesthetic and ethical dimensions of human experience. Such criticism is not altogether unjustified. Increasingly, however, economists are expanding their horizons and applying the laws of economics to the full spectrum of human activities.

The struggle to improve one's lot is not limited to the attainment of material goals. Although most economic principles have to do with the pursuit of material gain, they can be relevant to aesthetic and humanistic goals as well. The appreciation of a poem or play can be the subject of economic inquiry. Poems and plays, and the time in which to appreciate them, are also scarce.

Jacob Viner, an economist active in the first half of this century, once defined economics as what economists do. Today economists study an increasingly diverse array of topics. As always, they are involved in describing market processes, methods of trade, and commercial and industrial patterns. They also pay considerable attention to poverty and wealth; to racial, sexual, and religious discrimination; to politics and bureaucracy; to crime and criminal law; and to revolution. There is even an economics of group interaction, in which economic principles are applied to marital and family problems. And there is an economics of firm organization and the structure of incentives inside firms. Thus, although economists are still working on the conventional problems of inflation, unemployment, international monetary problems, and pricing policies, they are also studying the delivery of housing to the disadvantaged or of health care to the very young and the elderly. In one way or another, today's economists are tackling a wide variety of subjects, including committee structure, the criminal justice system, firm pay policies, ethics, voting rules, and the legislative process. Before this book and course have been completed, much will be said of how firms like General Electric, Microsoft, or Netscape can be expected to price their products, and we will touch on the conditions under which firms can be expected to give away their products (or even pay buyers to take their products). In fact, because we understand your professional goals for pursuing an MBA degree, we will never present theory for theory's sake. We will, in each and every chapter, show you how the theory can be used in practice by managers.

What is the unifying factor in these diverse inquiries? What ties them all together and distinguishes the economist's work from that of other social scientists? Economists take a distinctive approach to the study of human behavior. They employ a mode of analysis based on certain presuppositions about human behavior. For example, much economic analysis starts with the general proposition that people prefer more to fewer of those things they want and that they seek to maximize their welfare by making reasonable, consistent choices in the things they buy and sell. These propositions enable economists to derive the "law of demand" (people will buy more of any good at a lower price than at a higher price, and vice versa) and many other principles of human behavior.

One purpose of this book is to describe this special approach in considerable detail—to develop in precise terms the commonly accepted principles of economic

analysis and to demonstrate how they can be used to understand a variety of problems, including pollution, unemployment, crime, and ticket scalping. In every case, economic analysis is useful only if it is based on a sound theory that can be evaluated in terms of real-world experience.

Developing and Using Economic Theories

The real world of economics is staggeringly complex. Each day millions of people engage in innumerable transactions, only some of them involving money, and many of them undertaken for contradictory reasons. To make sense of all these activities, economists turn to theory.

A theory is a model of how the world is put together; it is an attempt to uncover some order in the seemingly random events of daily life. Economic theory is abstract, but not in the sense that its models lack concreteness. On the contrary, good models are laid out with great precision. Economic theories are simplified models *abstracted from* the complexity of the real world. Economists deliberately concentrate on just a few outstanding features of a problem in an effort to discover the laws that govern the relationships among them. Generally, a **theory** is a set of abstractions about the real world in which we work. An economic theory is a simplified explanation of how the economy or part of the economy functions or would function under specific conditions.

Quite often the economist must also make unproved assumptions, called simplifying assumptions, about the parts of the economy under study. For example, in examining the effects of price and availability on the amount of food sold, the economist might assume that people eat only oranges and bananas in the model society in question. Such a simplifying assumption is permissible in constructing a model, for two reasons. First, it makes the discussion more manageable. Second, it does not alter the problem under study or destroy its relevance to the real world.

As following chapters will reveal, economic theorizing is largely deductive—that is, the analysis proceeds from very general propositions (such as "more is preferred to less") to much more precise statements or predictions (for example, "the quantity purchased will rise when the price falls").[4] Economic theories sometimes vary in their premises and conclusions, but all develop through the following three steps.

First, a few very general premises or propositions are stated. "More is preferred to less" or "People will seek to maximize their welfare" are examples of such propositions. The premises tend to be so general that they are beyond dispute, at least to the economists developing the theory.

Second, logical deductions, which are tentative predictions about behavior, are drawn from the premises. From the premise "People will seek to maximize their welfare" we can deduce how people will tend to allocate their incomes at certain prices. We can then conclude that they will purchase more of a good when its price falls. Mathematics and graphic analysis are often very useful in deducing the consequences of premises.

---

[4] In contrast, inductive theorizing proceeds from very precise statements about observable relationships.

Third, the predictions are tested against observable experience. Theory may tell us that people buy more at lower prices than at higher prices, but the critical question is whether that prediction is borne out in the real world. Do people actually buy more apples when the price falls? Empirical tests require data to be carefully selected and statistically analyzed.

Empirical tests can never prove a theory's validity. The behavior that is observed—more apples purchased, for instance—may be caused by factors not considered in the theory. That is, the quantity of apples purchased may increase for some reason other than a drop in price. Empirical tests can only fail to disprove a theory. If a theory is repeatedly evaluated in different circumstances and is not disproven, however, its usefulness and general applicability increase. Economists have considerable confidence in the proposition that price and quantity purchased are inversely related because it has been repeatedly tested and found to be accurate.

Although a theory is not a complete and realistic description of the real world, a good theory should incorporate enough data to simulate real life. That is, it should provide some explanation for past experiences and permit reasonably accurate predictions of the future. When you evaluate a new theory, ask yourself: Does this theory explain what has been observed? Does it provide a better basis for prediction than other theories?

Positive and Normative Economics

Economic thinking is often divided into two categories—positive and normative. **Positive economics** is that branch of economic inquiry that is concerned with the world as it is rather than as it should be. It deals only with the consequences of changes in economic conditions or policies. A positive economist suspends questions of values when dealing with issues suck as crime or minimum wage laws. The object is to predict the effect of changes in the criminal code or the minimum wage rate—not to evaluate the fairness of such changes. **Normative economics** is that branch of economic inquiry that deals with value judgments—with what prices, production levels, incomes, and government policies *ought* to be. A normative economist does not shrink from the question of what the minimum wage rate ought to be. To arrive at an answer, the economist weighs the results of various minimum wage rates on the groups affected by them—the unemployed, employers, taxpayers, and so on. Then, on the basis of value judgments of the relative need or merit of each group, the normative economist recommends a specific minimum wage rate. Of course, values differ from one person to the next. In the analytical jump from recognizing the alternatives to prescribing a solution, scientific thinking gives way to ethical judgment.

Microeconomics and Macroeconomics

The discipline of economics is divided into two main parts—**microeconomics** and **macroeconomics**. As the term micro (as in microscope) suggests, **microeconomics** is the study of the individual markets—for corn, records, books, and so forth—that operate within the broad national economy. When economists measure, explain, and predict the demand for specific products such as bicycles and hand calculators, they are dealing with

microeconomics. Much of the work of economists is concerned with microeconomic analysis—that is, with the interpretation of events in the marketplace and of personal choices among products. This book, which has been designed with MBA students in mind, will deal almost exclusively with microeconomic theory, policy implications, and applications inside firms.

Questions of interest to microeconomists include:

What determines the price of particular goods and services?

What determines the output of particular firms and industries?

What determines the wages workers receive? The interest rates lenders receive? The profits businesses receive?

How do government policies—such as minimum wage laws, price controls, tariffs, and excise taxes—affect the price and output levels of individual markets?

Why do incentives matter inside firms and how can economic theory be used to properly structure a firm's incentives to increase worker productivity and firm profitability?

Economists are also interested in measuring, explaining, and predicting the performance of the economic system itself. To do so, they study broad subdivisions of the economy, such as the total output of all firms that produce goods and services. **Macroeconomics** is the study of the national economy as a whole or of its major components. It deals with the "big picture," not the details, of the nation's economic activity.

Instead of concentrating on how many bicycles or hand calculators are sold, macroeconomists watch how many good and services consumers purchase in total or how much money all producers spend on new plants and equipment. Instead of tracking the price of a particular good in a particular market, macroeconomics monitors the general price level or average of all prices. Instead of focusing on the wage rate and the number of people employed as plumbers or engineers, macroeconomists study incomes of all employees and the total number of people employed throughout the economy. In short, macroeconomics involves the study of national production, unemployment, and inflation. For that reason it is often referred to as aggregate economics.

Typical macroeconomic questions include:

What determines the general price level? The rate of inflation?

What determines national income and production levels?

What determines national employment and unemployment levels?

What effects do government monetary and budgetary policies have on the general price, income, production, employment, and unemployment levels?

These and similar questions are of more than academic interest. The theories that have been developed to answer them can be applied to problems and issues of the real world. They clearly have application to business, given that firm sales are often affected by "macro variables" such as national income and the inflation rate. Throughout this

book, as well as in specific chapters on topics such as regulation and deregulation, and price controls and consumer protection, we will examine the practical applications of economic theory.

However, we hasten to add that this book and course are devoted primarily to "microeconomic" theory and applications. We make microeconomics our focus because the issues at stake are more relevant to the interests of MBA students and because the microeconomic theory is generally viewed as being sounder than macroeconomic theory. Besides, we are firmly convinced that an understanding of the "macroeconomy" is necessarily dependent on an understanding of the "microeconomy."

In microeconomics we start with the proposition that all actions are constrained by the fact of scarcity. That is to say, in some basic way, scarcity—and the economic question of how to deal with it—touches all of us in how we do business and conduct our lives. We now turn to a study of property rights. Private "property rights" are one of the institutional mechanisms people have devised to help alleviate the pressing constraints of scarcity, which is why we take them up at this early stage in the course.

The Meaning and Importance of Property Rights

Property rights pertain to the permissible use of resources, goods, and services; they define the limits of social behavior and, in that way, determine what can be done by individuals in society. They also specify whether resources, goods, and services are to be used privately or collectively by the state or any smaller group.

Property rights are a social phenomenon; they arise out of the necessity for individuals to "get along" within a social space in which all wish to move and interact. Where individuals are isolated from one another by natural barriers or are located where goods and resources are abundant, property rights have no meaning. In the world of Robinson Crusoe, shipwrecked alone on an island, property rights were inconsequential. His behavior was restricted by the resources found on the island, the tools he was able to take from the ship, and his own ingenuity. He had a problem of efficiently allocating his time within these constraints—procuring food, building shelter, and plotting his escape; however, the notion of "property" did not restrict his behavior—it was not a barrier to what he could do. He was able to take from the shipwreck, with immunity, stores that he thought would be most useful to his purposes.[5]

After the arrival of Friday, the native whom Robinson Crusoe saved from cannibals, a problem of restricting and ordering interpersonal behavior immediately emerged. The problem was particularly acute for Crusoe because Friday, prior to coming to Tibago, was himself a cannibal. (Each had to clearly establish property rights to his body.) The system that they worked out was a simple one, not markedly different from

[5] The absence of human beings affected also his idea of what was useful. Crusoe, in going through the ship, came across a coffer of gold and silver coins: "Thou art not worth to me, no, not taking off the ground; one of these knives is worth all this heap [of gold]." At first, he evaluated the cost of taking the coins in terms of what he could take in their place and decided to leave them. But on second thought, perhaps taking into consideration the probability of being rescued, he took the coins with him! See *Robinson Crusoe* by Daniel Defoe.

that between Crusoe and "Dog." Crusoe essentially owned everything. Their relationship was that of master and servant, Crusoe dictating to Friday how the property was to be used.

The notion of property rights is broadly conceived by economists. Property rights are most often applied to discussions of real estate and personal property (bicycles, clothes, etc.); they are also applicable to what people can do with their minds, their ability to speak, how they wear their hair, and if and when they must wear their shoes.

In common speech, we frequently speak of someone owning this land, that house, or these bonds. This conventional style undoubtedly is economical from the viewpoint of quick communications, but it masks the variety and complexity of the ownership relationship. What is owned are *rights* to *use* resources, including one's body and mind, and these rights are always circumscribed, often by prohibition of certain actions. To "own land" usually means to have the right to till (or not to till) the soil, to mine the soil, to *offer* those rights for sale, etc., but not to have the right to throw soil at a passerby, to use it to change the course of a stream, or to force someone to buy it. What are owned are socially recognized rights of action.[6]

Property rights are not necessarily distributed equally, meaning that people do not always have the same rights to use the same resources. Students may have the right to use their voices (i.e., a resource) to speak with friends in casual conversation in the hallways of classroom buildings, but they do not, generally speaking, have the right to disrupt an English class with a harangue on their political views. However, the English professor, although his behavior is circumscribed, has the right to "allow" his or her political views to filter into the English lectures. And if the President of the United States walked into the same English class and began speaking extemporaneously on *his* (or *her*) political views, it is not likely that anyone would object. A person has the right to go without shoes on a beach, but one does not *always* have the right to enter a restaurant without shoes. On the other hand, the restaurant owner's best friend may have that right. By the same token, although undergraduate students generally pay a fraction of their educational expenses at state universities, they have the right to university facilities such as tennis courts and the university bookstore, but nonstudent taxpayers do not have the same rights to these facilities.

In other words, property rights can be recast in terms of the *behavioral rules*, which effectively limit and restrict our behavior. Behavioral rules determine what rights we have with regard to the use of resources, goods, and services. The rights we have may be the product of the legislative process and may be enforced by a third party: usually the third party is the government or, more properly, the agents of government. In this case property rights emerge from laws.

On the other hand, rules that establish rights may not have third-party enforcement. In this case they carry weight in the decisions of individuals simply because individuals recognize and respect behavioral limits for themselves and others. They may do this because of the value they attach to "living up" to their contractual agreements, which may be implied in their associations with others, and because their

---

[6] Armen A. Alchian and Harold Demsetz, "The Property Rights Paradigm," *Journal of Economic History*, vol. 33, p. 17, March 1973.

own rights may be violated if they violate the rights of others. Two neighbors may implicitly agree to certain modes of behavior, such as not mowing their lawns on Sunday mornings or playing their stereo equipment late at night.[7] Their behavior may be in recognition of what it means to be a "good neighbor" and of what life can be like if limits to their behavior are not observed. The neighbor who starts his mower early Sunday morning may hear music late at night or may find his rights invaded in other ways. More will be said on this, but for now we mean only to point out that the behavior of each through "offensive and counteroffensive" maneuvers may deteriorate into a state in which both parties are worse off than they would have been if restrictions on their own behavior were commonly observed. From this we see the bases for *behavioral rules* or, what amounts to the same thing, *property rights.*

Property rights are important to any inquiry of social order because it is on the basis of such rights that the terms *individual* and *state* are given social meaning, that actions are delimited, and that a specific social order will emerge. The existing property rights structure is predicated upon specific social and physical conditions. Changes in those conditions can cause a readjustment in the nature of social order.

Property Rights and the Market

In the private market economy people are permitted to initiate trades with one another. Indeed, when people trade, they are actually trading "rights" to goods and services or to do certain things. For example, when a person buys a house in the market, he is actually buying the right to live in the house under certain conditions, for example, as long as he does not disturb others. This market economy is predicated upon establishing patterns of *private* property rights; those patterns have legitimacy because of enforcement by government and, perhaps just as important, because of certain precepts regarding the limits of individual behavior that are commonly accepted and observed.[8] Without recognized property rights there would be nothing to trade—no market.

How dependent are markets on government enforcement for the protection and legitimacy of private property rights? Our answer must of necessity be somewhat speculative. We know that markets existed in the "Old West" when *formally* instituted governments were nonexistent. Further, it is highly improbable that any government can be so pervasive in the affairs of people that it can be the arbiter of all private rights. Cases in which disputes over property rights within college dormitories are settled by student councils are relatively rare, and the disputes that end up in the dean's office or at police headquarters are rarer still. Most conflicts over property rights are resolved at a local level, between two people, and many potential disputes do not even arise because of generally accepted behavioral limits.

Finally, the concept of property rights helps make clear the relationship between the public and private sectors of the economy—that is, between that section of the

---

[7] This is an example used by James M. Buchanan, *The Limits of Liberty* (Chicago: University of Chicago Press, 1975), p. 20.

[8] In addition, there is considerable *private* enforcement of property rights. Almost all people take some measures to secure their own property. They put locks on their doors, leave lights on at night, and alert their neighbors to take their newspapers in when they are out of town.

economy organized by collective action through government and that section which is organized through the actions of independent individuals. When government regulates aspects of the market, it redefines behavioral limits (in the sense that people can no longer do what they once could) and can be thought of as realigning the property rights between the private and public spheres. When the government imposes price ceilings on goods and services, as it did during the summer of 1971, it is redefining the rights that sellers have with regard to the property they sell. One of the purposes of economics is to analyze the effect that a realignment of property rights has on the efficiency of production.

Anarchy: A State of Disorder

Property rights are so much a part of our everyday experience that we are inclined to think of them as being "natural," a part of our birthright. The Declaration of Independence speaks of "certain unalienable rights." Indeed, it is hard to imagine a world in which people interact within a defined social space without the existence of property rights. The purpose of this section is to envision such a state in order to gain some insight into the origins of property rights and, therefore, social order.

Thomas Hobbes, a seventeenth-century political scientist philosopher, envisioned a state in which there was a complete absence of property rights, either those rights that have legitimacy because of their social acceptability or those that exist because of legal enforcement. He called this "the state of nature," and his analysis was not very attractive. Because Hobbes gave very little credence to social acceptance as a basis for property rights, his attention was on the role of the state. He believed that "during the time men live without a common Power to keep them in awe," every man will be pitted against another in continual struggle for dominance and protection. Life will be "solitary, poore, nasty, brutish, and short." Where there is no state, he argued, there will be no law and therefore, "no Property ... no *Mine* and *Thine* distinct, but only that to be every man's that he can get, and for so long as he can keep it."[9]

One of Hobbes' purposes in writing *Leviathan* was to justify the sovereign state as an absolutely necessary political entity. He tried to convince his contemporaries of the potential for conflict among men without the state; that it is necessary to hand over considerable political power to the state in order that internal conflicts may be minimized. He argued that it is in man's self-interest to swear full allegiance to the state.

In order to make his argument as convincing as possible, it was somewhat natural for Hobbes to describe "the state of nature" in the worst possible terms. One can accept the criticism that Hobbes exaggerated the need for the state without ignoring a cornerstone of his argument: Without legally defined property rights, there is considerable potential for conflict among men. The life of man in the state of nature may not invariably be "solitary, poore, nasty, brutish, and short," but it may be markedly less comfortable without property rights than with them.

---

[9] Thomas Hobbes, *Leviathan*, ed. By C.B. Macpherson [Baltimore, Md.: Penguin Books, Inc., 1968 (first published 1651], pp. 185–88.

In an idealized world in which people are fully considerate of each other's feelings and adjust and readjust their behavior to that of others without recourse to anything resembling a dividing line between "mine" and "thine," property rights are no more necessary than they were for Robinson Crusoe alone on Tibago. But in the world as it now exists, there is the potential for conflict. Granted, the potential may not be present in all our interpersonal experiences. People have interests that, for all practical purposes, are independent of one another, and many of our interests are perfectly congruent with the interests of those around us. However, people have spheres of interests (described for two people by the circle in Figure 1.1) that extend outward from themselves and that intersect with the interests of others. A basic axiom of behavior (one to be developed in greater detail later) is that most people want more than they have, which means they have an interest in, or can benefit from, that which others have. In other words, they have competing interests—or, in terms of Figure 1.1, areas where their spheres of interests intersect. It is here that the potential for conflict arises, that a dividing line between "mine" and "thine" must be drawn.

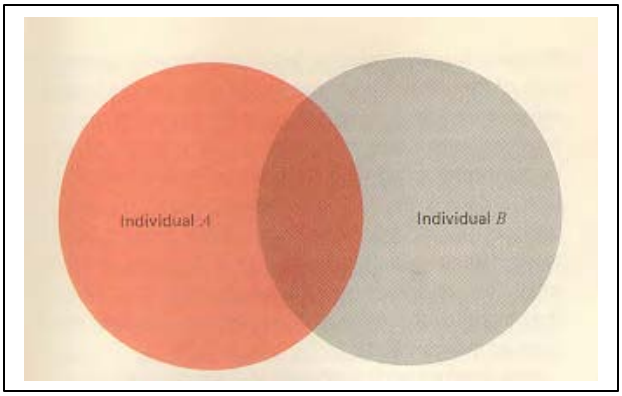


Figure 1.1 Individuals have spheres of interest, which are illustrated, by the two circles. The intersection of the two circles represents the arc of potential conflict between two individuals; it is the area within which property rights (or behavioral limits) must be established.

Children at play provide us with a reasonably clear illustration of the absence of and potential for conflict among people in the larger community. Children can often play together for long periods of time without conflict. They each have interests that do not invade the interests of others (which may be described by the clear portions of the circles in Figure 1.1); for example, one may want to play with a truck, one with a bucket and shovel, and another with toy cowboys. For periods, their behavior may approximate the idealized society mentioned above. On the other hand, when two children want to play with the same toy or play the same role of mother or father in their game of "house," or when one child wants to take over the entire sandbox, conflict is revealed, first with harsh words, possibly in fights, leading to a breakdown of their social interaction—play.

Conflict or the potential for conflict can be alleviated by the development of property rights, held either communally, by the state, or by private individuals. These rights can be established in ways that are similar but which can be conceptually distinguished: (1) *voluntary* acceptance of behavioral norms with no third-party enforcer, such as the police and courts, and (2) the specification of rights in a legally binding "social contract," meaning that a third-party enforcer is established. Most of what we say for the remainder of this chapter applies to both modes of establishing rights. However, for reasons developed later in the book, the establishment of rights through voluntary

acceptance of behavioral norms, although important in itself, has distinct limitations, especially in relation to size. More specifically, many behavioral norms have a tendency to break down in large-group settings. Because most people hold to the behavioral norm that they should not pollute, and yet at least to some degree they pollute anyway, and because legal codes are filled with specifications of property rights, meaning something has failed, the limitations of behavioral norms may come as no surprise. Be that as it is, holding the discussion of voluntary behavioral rules until later in the book will permit us to narrow our attention and, perhaps, gain a deeper understanding of the basis of legal property rights. For now, let's step back and consider in more detail the social basis for property rights.

The Emergence of Property Rights

To develop the analysis in the simplest terms possible, consider a model of two people, Fred and Harry, who live alone on an island. They have, at the start, no behavioral rules or anything else that "naturally" divides their spheres of interest. That is, they have nothing that resembles property rights. Further, being rational, they are assumed to want more than they can produce by themselves. Their social order is essentially anarchic. Each has two fundamental options for increasing his welfare: He can use his labor and other resources to produce goods and services or he can steal from his fellow man. With no social or ethical barriers restricting their behavior, they should be expected to allocate their resources between these options in the most productive way. This may mean that each should steal from the other as long as more is gained that way than through the production of goods and services.

If Fred and Harry find stealing a reasonable course to take, each will have to divert resources into protecting that which he has produced (*or* stolen). Presumably, their attacks and counterattacks will lead them toward a social equilibrium in which each is applying resources to predation and defense and neither finds any further movement of resources into those lines of activity profitable.[10] This is not equilibrium in the sense that the state of affairs is a desirable one; in fact, it may be characterized as a "Hobbesean jungle" in which "every man is Enemy to every man."

In an economic sense, the resources diverted into predatory and defensive behavior are wasted; they are taken away from productive processes. If these resources are applied to production, total production can rise, and both Fred and Harry can be better off—both can have more than if they try to steal from each other. It is only through winding up in a state of anarchy or seeing the potential for ending up there that they must question the rationality of continued plundering and unrestricted behavior; and it is because of the prospects of individual improvement that there exists a potential for a "social contract" that spells out legally defined property rights. Through a social contract they may agree to place restrictions on their own behavior, but they will do away with the restraints that, through predation and required defense, each imposes on the other. The fear of being attacked on the streets at night can be far more confining than laws that

[10] For a rather difficult discussion of "equilibrium" under anarchy, see Winston C. Bush, "Individual Welfare in Anarchy," in Gordon Tullock (ed.), *Explorations in the Theory of Anarchy* (Blacksburg, Va.: University Publications, Inc., 1972), pp. 5–8.

restrict people from attacking one another. This is what John Locke meant when he wrote, "The end of law is not to abolish or restrain but to preserve and enlarge freedom."[11]

Once the benefits from the social contract are recognized, there *may* still be, as in the case of voluntary behavioral norms, an incentive for Fred or Harry to chisel on the contract. Fred may find that although he is "better off" materially by agreeing to property rights than he is by remaining in a state of anarchy, he may be even "better off" by violating the agreed-upon rights of the other. Through stealing, or in other ways violating Harry's rights, Fred can redistribute the total wealth of the community toward himself.

To illustrate, consider Figure 1.2, which contains a chart or matrix of Fred and Harry's utility (or satisfaction) levels if either respects or fails to respect the rights established for each as a part of the contract. (The actual utility levels are hypothetical but serve the purpose of illustrating a basic point.) There are four cells in the matrix, representing the four combinations of actions that Fred and Harry can take. They can both respect the agreed-upon rights of the other (cell 1), or they can both violate each other's rights (cell 4). Alternatively, Harry can respect Fred's rights while Fred violates Harry's rights (cell 3), or vice versa (cell 2).

Clearly, by the utility levels indicated in cells 1 and 4, Fred and Harry are both better off by respecting each other's rights than by violating them. However, if Harry respects Fred's rights and Fred fails to reciprocate, Fred has a utility level of 18 utils, which is greater than he will receive in cell 1, that is, by going along with Harry and respecting the other's rights. Harry is similarly better off if he violates Fred's rights while Fred respects Harry's rights: Harry has a utility level of 16, whereas he will have a utility level of 10 utils if he and Fred respect each other's rights. The lesson to be learned: Inherent in an agreement over property rights is the possibility for each person to gain by violating the rights of the other. If both follow this course, they both will end up in cell 4, that is, back in the state of anarchy.

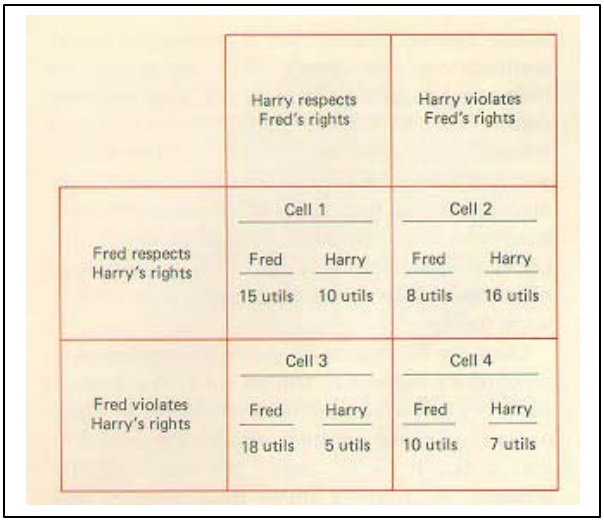


Figure 1.2 The payoffs (measured in "util" terms) from Fred and/or Harry either respecting or violating the other's rights are indicated in the four cells of the matrix. Each has an incentive to violate the other's rights. If they do violate each other's rights, they will end up in cell 4, the worst of all possible states for both of them. The productivity of the "social contract" can be measured by the increase in Fred and Harry's utility resulting from their moving from cell 4, the "state of nature," to cell 1, a state in which a social contract is agreed upon.

[11] Locke, *The Second Treatise*, p. 32.

There are two reasons why this may happen. First, as we stated above, both Fred and Harry may violate each other's rights in order to improve their own positions; the action may be strictly *offensive.* By the same token, each must consider what the other will do. Neither would want to be caught upholding the agreement while the other one violates it. If Fred thinks Harry may violate his rights, Fred may follow suit and violate Harry's rights: he will be better off in cell 4, i.e., anarchy, than in cell 2. Fred and Harry can wind up in anarchy for purely *defensive* reasons. Many wars and battles, both at the street and international levels, have been fought because one party was afraid the other would attack first in order to get the upper hand. The same problem is basically involved in our analysis of the fragile nature of Fred and Harry's social contract.

Fred and Harry's situation is a classic example of what social scientists call a "prisoner's dilemma." The name comes from a standard technique of interrogation employed by police to obtain confessions from two or more suspected partners to a crime. If the method is used, the suspects are taken to different rooms for questioning, and each is offered a lighter sentence if he confesses. But each will also be warned that if the other suspect confesses and he does not, his sentence will be more stringent. The suspect has to try to figure out, without the benefit of communication, how the other will stand up to that kind of pressure. Each may worry that the other will confess and may confess because he cannot trust his partner not to take the easy way out.[12] The problem that the individual suspect becomes more complicated when the larger the number of partners to the crime who are caught with the individual increases. There are more people upon whom he must count to hold up under the pressure, which he knows is being brought to bear. He must also consider the fact that the others may confess because they cannot count on all partners to hold under the pressure.

To prevent violations, both of offensive and of defensive nature, a community may agree to the establishment of a police, court, and penal system to protect the rights specified in the social contract. The system may be costly, but the drain on its total wealth may be smaller than if it reverts back to anarchy, in which case resources will be diverted into predatory and defensive behavior. The costs associated with making the contract and enforcing it will determine just how extensive the contract will be, and this matter will be considered later in a separate chapter; that for now, assuming the benefits from the contract exceed the costs of contracting and enforcement, we may summarize the foregoing discussion in terms of Figure 1.3. In the state of nature, Fred and Harry, through allocating their resources among productive, predatory, and defensive uses, will achieve a certain level of welfare. In terms of Figure 1.3, Fred achieves an initial utility level of U_{F1} and Harry, U_{H1}. By developing a social contract, through which they define and enforce property rights, each can move to a higher utility level; Fred to U_{F2} and Harry to U_{H2}. With social contracts, they both can move to higher utility levels because they no longer have to divert their resources to predatory and defensive actions.

[12] There is no wonder that prisoners have such harsh feelings toward those who cave in and "rat on them" or "fink out."

The Emergence of Exchange

The social contract, which defines property rights, establishes only the limits of permissible behavior; it does not mean that Fred and Harry will be satisfied with the property rights they have been given through the contract. To the degree that some other combination will give them more satisfaction, exchanges can emerge, provided, of course, that the social contract permits them. In terms of Figure 1.3, they can, through exchanges or trades, increase their utility to U_{F3} and U_{H3}.

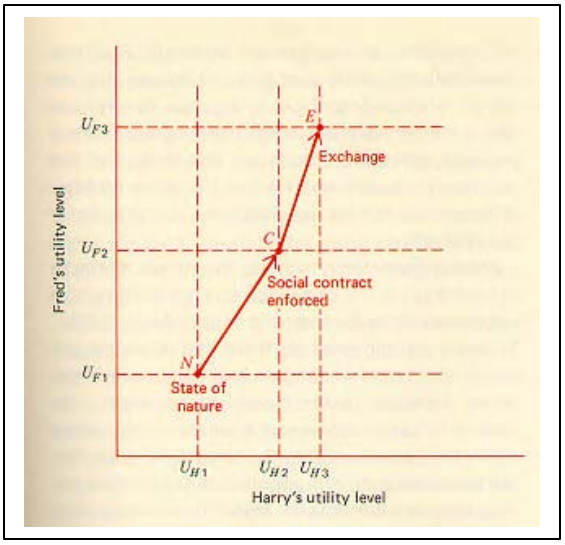


Figure 1.3 In the "state of nature," in which Fred and Harry each has to use resources to fend off the other, the welfare of Fred and Harry are, respectively, UF1 and UH1, or point *N*. A social contract can move them both to point *C*. They can further improve their welfare by trading the "rights" to goods and services that they are given in the social contract.

For example, suppose that the only goods on Fred and Harry's island are coconuts and papayas. The social contract specifies the division of the fruits between them. We need not concern ourselves with the total number of the fruit each has; we need only indicate the relative satisfaction that Fred and Harry receive from the marginal units. Suppose the marginal utilities in the table below represent the satisfaction they received from the last coconut and papaya in their possession:

	Coconut	Papaya
Fred	10 utils	15 utils
Harry	90 utils	30 utils

In the illustration, Fred receives more utility from the last papaya (15 utils) than from the last coconut (10 utils). He would be on a higher level of utility if he could trade a coconut for a papaya. He would lose 10 utils from the coconut but would more than regain that with the additional papaya. On the other hand, Harry receives more utility from the last coconut than from the last papaya. He would gladly give up a papaya for a coconut; he would be 60 utils of satisfaction better off (90 minus 30) than if he did not

engage in the exchange. The two should continue to exchange *rights* to the coconuts and papayas until one or both of them can no longer gain via trade.

In this example, we are not concerned with production of coconuts and papayas; we are concerned merely with the benefits from trade resulting from the initial allotments of the fruits. The trades are comparable to those that took place in the prisoner-of-war camps as described by R.A. Radford. (See the first few pages of this text.) If the social contract allocates to Fred and Harry rights to *produce* the fruit, we can also demonstrate that both can be better off through specializing in their production and trading with each other. Consider the information in the following table; it indicates how many coconuts or papayas Fred and Harry can produce with, say, one hour of labor:

	Coconut Production	Papaya Production
Fred	4	8
Harry	6	24

In one hour of labor Fred can produce either 4 coconuts or 8 papayas; Harry can produce either 6 coconuts or 24 papayas. Even though Harry is more productive in both lines of work, we can show that they both can gain by specializing and trading with each other.

If Fred produces 4 coconuts, he cannot use that hour of time to produce the 8 papayas. In other words, the cost of the 4 coconuts is 8 papayas, or, what amounts to the same thing, the cost of 1 coconut is 2 papayas. Fred would be better off if he could trade 1 coconut for *more than* 2 papayas, because that is what he has to give up in order to produce the coconut. To determine whether there is a basis for trade, we must explore the cost of coconuts and papayas to Harry. We note that the cost of 1 coconut to Harry is 4 papayas; this is because he has to give up 24 papayas to produce 6 coconuts. If Harry could give up less than 4 papayas for a coconut, he would be better off. He could produce the 4 papayas; and if he has to give up fewer than that for a coconut, he will have papayas left over to eat, which he would not have had without the opportunity to trade.

To summarize: Fred would be better off if he could get more than 2 papayas for a coconut; Harry would be better off if he could give up fewer than 4 papayas for a coconut. If, for example, they agree to trade at the exchange rate of 1 coconut for 3 papayas, both would be better off. Fred will produce a coconut, giving up 2 papayas, but he can turn around and get 3 papayas for the coconut. Hence, he is better off. Harry can produce 4 papayas, giving up 1 coconut, and trade 3 of the papayas for a coconut. He has the same number of coconuts, but has an additional papaya. Harry is better off.

Although relatively simple, the above example of exchange is one of economists' most important contributions to discussions of social interaction. So many people seem to think that when people trade, one person must gain at the expense of another. If people in the United States trade with people in Japan, someone must be made worse off in the process, or so the argument goes. We will deal with such arguments in more detail

in the last chapter in the book on international trade; for now we wish to emphasize that we have demonstrated that, through trade, both Harry and Fred are better off. This was demonstrated even though we postulated that Harry was more efficient than Fred in the production of both fruits!

Communal Property Rights

To many, the ideal state of affairs may appear to be one in which everyone has the right to use all resources, goods, and services and in which no one (not even the state) has the right to exclude anyone else from their use. We may designate such rights as "communal rights." Many rights to scarce property have been and still are allocated in this way. Rights to the use of a university's facilities are held communally by the students. No one admitted to the university has the right to keep you off campus paths or lawns or from using the library according to certain rules and regulations. (Such rules and regulations form the boundaries, much as if they were natural, within which the rights are truly communal.) The rights to city parks, sidewalks, and streets are held communally. Before our country was settled, many Indian tribes held communal rights to hunting grounds: that is, at least within the tribe's territory, no one had the right to exclude anyone else from hunting on the land. During most of the first half of the nineteenth century, the rights to graze cattle on the prairies of the western United States were held communally; anyone who wanted to let his cattle loose on the plains could do so. Granted, the United States government held by law the right to exclude people from the plains; but as long as it did not exercise that right, the land rights were communal. The same can be said for all other resources whose "owner" does not exercise the right to exclude.

Communal property rights can be employed with tolerably efficient results so long as one of two conditions holds: (1) there is more of the resource than can be effectively used for all intended purposes (in other words, there is no cost to its use) or (2) people within the community fully account for the effects that their own use of the resources has on others. Without the presence of one of these conditions, the resources will tend to be "overused."

Under communal ownership, if the resource is not presently being used by someone else, no one can be excluded from the use of it. Consequently, once in use, the resource becomes, for that period of time, the private property of the user. The people who drive their cars onto the freeway take up space on the road that is not in use; no one else (they hope!) can then use that space at the same time. Unless the drivers violate the rules of the road, they cannot be excluded from that space; and if they are rational, they will continue to use the resource until the marginal cost of doing so equals the marginal benefits to them. They may consider most of the costs involved in their use of the road, but one that they may overlook, especially as it applies to themselves personally, is that their space may have had some alternative use: that is, by others.[13] Their presence also increases highway congestion and the discomfort of the other drivers. As a result, they may overextend the use of their resource, meaning they continue to drive as long as the additional benefits *they*, themselves, get from driving additional miles is greater than the

[13] Environmentalists argue that many roads should not have been built; the alternative use in this case would be scenery, for example.

additional cost. However, they can overlook the cost they impose on others, which can mean that the *total* cost for everyone driving additional miles is greater than the total benefits.

The state can make the driver consider the social costs of driving in an indirect way by imposing a tax on the driver's use of the road equal to the distance between *a* and *b*. This is called "internalizing the social cost." Once the state does this—and it is commonly done through gasoline taxes and/or tolls—the rights to the freeway are no longer "communal"; the rights have been effectively taken over by the state.

There are two additional ways that social costs can be internalized. First, people can be considerate of others and account for the social cost in their behavior. Second, the right to the road can be turned into *private property*, meaning that individuals are given the right to exclude others from the use of the resource (i.e., the road). This may seem to be a totally undesirable turn of events unless we recognize that private owners can then charge for the use of the road: they can sell "use rights," in which case the marginal cost of driving will rise, resulting in an increase in the cost that individual drivers incur.

The prime difference between this private ownership and government taxation is that with private ownership, the revenues collected go into the coffers of individuals instead of to the state; this is either "good" or "bad," depending upon your attitude toward government versus private uses of the funds. Furthermore, under private ownership and without viable competitors (and we have an example in which competition *may* not be practical), the owners may attempt to charge an amount that is greater than the social costs in the figure; they may attempt, in the jargon of economists, to acquire *monopoly profits*, and in so doing cause an *underuse* of the road.[14] (A monopoly is a single seller of a good or service that can charge higher prices and reap greater profits than if it had to worry about the actions of other competitors.)

For that matter, the state-imposed taxes may be greater than the social costs. The state may also act like a monopolist. State agencies may not be permitted to make a "profit" as it is normally conceived, but this does not exclude the use of their revenues for improving salaries and the working conditions of state employees. Monopoly profits may be easy to see on the accounting statements of a firm but may be lost in bureaucratic waste or over-expenditures under state ownership. State ownership does not necessarily lead to waste, but it is a prospect, and one only that the naïve will ignore. More is said on the subject at various points in the book.

We have now considered the distinction between private and communal property. Several examples will enable us to amplify that distinction and to understand more clearly the limitations of communal property rights and the pervasive use of private property.

---

[14] To provide for competition and to prevent monopoly profits from emerging, private rights can be assigned to similar units of the same resource. Although this may not be practical in road construction, it is quite practical in the cattle business, for example. Many different people can own all the resources necessary for cattle production. If one tries to raise his price to achieve monopoly profits, the others can undercut him, forcing him to lower his price. As a general rule, competition requires the dispersion of property rights among different people and groups.

Pollution

Pollution can be described as a logical consequence of communal property rights to streams, rivers, air, etc. The state and federal governments, by right of eminent domain, have always held rights to these resources; but until very recently they have inadequately asserted their right to exclude people and firms from their use. As a result, the resources have been subject to communal use and to overuse, in the same sense as that discussed above.

By dumping waste into the rivers, people, firms, and local governments have been able to acquire ownership to portions of the communal resource—they use it and pollute it. Furthermore, because of the absence of exclusion, those people doing the polluting do not have to pay to draw the resource away from its alternative uses (such as pretty scenery) or to reimburse the people harmed by the pollution for the damage done. Under communal ownership, in which government does not exercise its control, the firm with smoke billowing from its stacks does not have to compensate the people who live around the plant for the eye irritation they experience or the extra number of times they have to paint their homes.

Pollution is often thought to be the product of antisocial behavior, as indeed it often is. Many who pollute simply do not care about what they do to others. However, much pollution results from the behavior of people who do *not* have devious motives. People may view their behavior as having an inconsequential effect on the environment. The person who throws a cigarette butt on the ground may reason that if this cigarette butt is the only one on the ground, it will not materially affect anyone's sensibilities, and in fact it may not. However, if everyone follows the same line of reasoning, the cigarette butts will accumulate and an eyesore will develop. Even then, there may be little incentive for people to stop throwing their butts on the ground. Again, a person may reason on the basis of the effects of his own individual action: "If I do not throw my butt on the ground here with all the others, will my behavior materially affect the environment quality, *given the fact that other butts are already there?*" This type of reasoning can lead to a very powerful argument for conversion of communal rights to private or state rights, with the implied power for someone to exclude some or all of this kind of use.[15]

Fur Trade

According to Harold Demsetz, the hunting grounds of the Indian tribes of the Labrador Peninsula were held in common until the emergence of the fur trade there.[16] The Indians could hunt as they wished without being excluded by other members of their tribe. Presumably, given the cost of hunting and the limited demand for meat, there was no inclination to "over-hunt," that is, until there was an adverse effect on the stock of animals in the area.

---

[15] For a similar discussion of why university campuses have dirt paths on the campus lawns, see Richard B. McKenzie and Gordon Tullock, *The New World of Economics* (Homewood, Ill.: Richard D. Irwin, Inc., 1978), chap. 2.

[16] Harold Demsetz, "Toward a Theory of Property Rights," *American Economic Review*, vol. 57, pp. 347–59, May 1964. Demsetz cites Eleanor Leacock, "The Montagnes 'Hunting Territory' and the Fur Trade," vol. 56, no. 5, part 2, Memoir No. 78.

However, when fur trading commenced and the Indians hunted animals for their skins, the demand and therefore the price of animal skins increased. This provided an incentive for the Indians to hunt beyond their demand for meat. Under communal ownership, when a beaver was killed, an Indian hunter did not have to consider the effects that his action had on the ability of the other hunters to trap and hunt. Each hunter, through his own efforts, imposed a cost on the others; when a beaver was killed by one hunter, the task of finding beavers was made more difficult for the other hunters. The cost may be construed as a social cost, much like the congestion a driver can impose upon the other drivers around. Furthermore, under a communal rights structure, there was little incentive for hunters to avoid trapping or incurring the costs of increasing the stock of animals. If a hunter refrained from killing a beaver, perhaps someone else would kill it. In addition, if one person tried to increase the stock of animals, perhaps many others would benefit from his efforts in terms of more animals for them to kill. There was, in other words, no assurance that the Indian who built up the stock of animals would reap the benefits. (For the same reason, we doubt that many buildings would be built if the developers could not reap the benefits of their investment or if what they built would be *communal* property upon completion.) The Indians' solution to the problem of overkill was to assign private property rights to portions of the hunting grounds. Each individual, by virtue of his right to exclude others, had an incentive to control his own take from the land and to take measures, much as ranchers do, to increase the potential stock of furs.

Whales and Seals

Whales have been hunted for centuries, but there has never been a problem with their possible extinction until the last two centuries. Whales have always been more or less communal property; however, because people in bygone centuries did not have the technology we now have to kill and slaughter whales far at sea, the sheer cost of hunting them prevented men from exceeding the whales' reproductive capacity. Theoretically, the problem could be solved by applying the same solution to the whale overkill as the Indians applied in their hunting grounds: establish private property rights. However, whales present a special problem. The annual migrations of whales can take them through 6,000 miles of ocean. Establishing and enforcing private property rights to such an expanse of ocean is an ominous task, even without the complications involved in securing agreement among several governments to respect those rights. These costs have, without doubt, been a major reason that whales remain communal property and are threatened still with extinction.

Communal property rights can also have an effect on the way animals are treated and slaughtered. Armen Alchian and Harold Demsetz have provided us with a vivid description of the seal slaughter in Canada:

> In 1970, the newspapers carried stories of the barbaric and cruel annual slaughter of baby seals on the ice floes off Prince Edward Island in the Gulf of St. Lawrence. The Canadian government permitted no more than 50,000 animals to be taken, so hunters worked with speed to make their kills before the legal maximum was reached. They swarmed over the ice

> floes and crushed the babies' skulls with heavy clubs. Government offices received many protests that the seals were inhumanly clubbed (by humans) and often skinned alive. The minister of fisheries warned the hunters of the strong pressure he was under to ban the hunt and that he would do so unless the killing methods were humane.[17]

Alchian and Demsetz point out that the Canadian government had effectively made the first 50,000 baby seals communal property among the hunters; the seals became private property when and only when they were killed. Possession of only the *first* 50,000 baby seals was legal. The rights to the seals were allocated communally on a first-come, first-served basis, and Alchian and Demsetz stressed that such a rationing system tends to encourage "rapid hunting techniques and to make a condition of their success the degree to which the hunter can be ruthless."[18]

MANAGER'S CORNER I: **How Incentives Count in Economics and Business**

We noted above that much of this book and course is concerned with the problem of overcoming a basic condition of life: *scarcity*. Firms are an integral means by which the pressures of scarcity are partially relieved for all those people who either own or work for or with the firms. However, in order to get people involved in or with firms to work diligently for the firms, they must have some reason or purpose—some *incentive*—to do that which they are supposed to do. Within sections of this book that we have titled "Manager's Corner," we seek to apply the economic principles developed in the first part of the chapter to problems that all MBA students will confront in their "real world" careers, that of getting incentives within firms right. Doing that is no easy assignment for managers mainly because incentives are powerful—both when they are wrong as well as right, as we will see by taking up an array of incentive issues that range from how workers' compensation can affect firm output to how a firm's finances (debt and equity) can affect management risk taking and, hence, firm profitability.

Incentives are growing in importance as a tool of management for several important reasons:

- Production of goods and services in many industries has become unbelievably sophisticated and complex, which has required managers to draw on the creativity, skills, and human capital of line workers who often have local information about their work—what can and cannot be done—that is not, and cannot be, available to their supervising managers.
- Production processes for many goods and services have become global in scope, which necessarily means many workers must work far removed from their supervisors, who have no way of monitoring what the workers are doing on a daily basis.

---

[17] Armen A. Alchian and Harold Demsetz, "The Property Rights Paradigm," *Journal of Economic History*, vol. 33, p. 20, March 1973.
[18] Ibid.

- Firms have, to a growing extent, relied on "outsourcing," which means firms are buying more and more of their inputs—from parts to human resource services—from outside suppliers whose business goals are not in line always with the business goals of the buyers.
- The hierarchical organizational structures of many firms have been "flattened," which implies fewer layers of managers and supervisors.
- Moreover, the pace of technological and organizational innovations and change has speeded up, increasing the extent to which decision making has been devolved to lower and lower levels within firms' organizational structures.

These ongoing, far-reaching changes in the economy have long been documented. What has not been fully appreciated is the fact that these changes mean that a growing number of workers must work apart from the direct supervision of their bosses. Because managers are less able to directly monitor the workers under them, old command-and-control methods of management have begun to wane. Managers have become less able to tell their employees what to do simply because the managers, in highly sophisticated and complex production processes, don't have the skills and knowledge to do what their employees can do or even figure out exactly what their workers should do with a substantial share of their time. Production has truly become "participatory," which means that higher managers must rely on their underlings to do what they are supposed to do.

Under the circumstances, managers must find ways to entice workers to use their creativity, skills, and human capital to pursue firm goals. In short, managers must use *incentives*. They can no longer manage by commands, at least not to the extent that they once could. They must now manage through incentives, which they are doing in growing numbers. The count of firms that tie manager and worker incomes to performance is not known, but few doubt that it is growing rapidly.[19] We submit that incentives are a popular solution for today's management dilemmas for a simple reason: *Incentives work and always have, often with dramatic effect.*

Incentives at Work

In the late nineteenth century, British boat captains were paid to carry prisoners from England to the wilds of Australia, just to rid England of its crime problem and reduce the cost of housing criminals. The captains were paid a flat fee for each prisoner who boarded at an English port, giving the captains had a strong incentive to board as many

[19] In 1945, there were only 2,113 firms in the United States that had deferred compensation or profit-sharing plans for their workforces. In 1991, the count of firms with such group incentive plans had risen to nearly 500,000 (as reported in Haig R. Nalbantian and Andrew Schotter, "Productivity Under Group Incentives: An Experimental Study," *American Economic Review*, vol. 87 (no. 3), June 1997, pp. 314–41). One researcher predicted in the late 1980s that by the turn of the century, a quarter of all firms listed on the American, New York, and over-the-counter Stock Exchanges will, because of distribution of shares of stock and stock options to workers, have more than 15 percent of their shares owned by their workers [Joseph R. Blasi, *Employee Ownership: Revolution or Ripoff?* (Cambridge, Mass.: Ballinger Publishers, 1988)].

prisoners as they could, but only a weak incentive to deliver them to Australia alive. In other words, the incentive system was perverse. If prisoners died along the way from lack of food and care, the cost of the trips was lowered. And the survival rate was a miserable 40 percent, a fact that outraged humanitarians then, as it would now! But despite the moral outrage at the time, the survival rate of the prisoners didn't budge until the incentive was changed. Edwin Chadwich, the government official in charge of the deportation of criminals in the 1860s, had a bright idea for restructuring the incentive system: pay the captains not by the count of prisoners who boarded the boats in England, but by the number of prisoners who disembarked in Australia. The survival rate rose quickly and dramatically to 98.5 percent! All because the captains then had a strong incentive to take care of their charges.[20]

Under the former Soviet Union, there was more than an ounce of truth in the widely circulated Soviet witticism: "We pretend to work, and they pretend to pay us." In the new economy, the pretense of work will not be rewarded. Former (and last) premier of the Soviet Union Mikhail Gorbachev made much the same point when he wrote that "amazing things happen when people take responsibility for everything themselves. The results are quite different, and at times people are unrecognizable. Work changes and attitudes to it, too."[21] Many world leaders worry that the Soviet people have become accustomed to being communists and will not make the transition to market thinking without grave difficulty. Further, they worry that instituting property rights and the attendant incentives may not have all the beneficial effects that they have had in the West. After all, the Soviet citizens need to rebuild their economy, and the rebuilding process has imposed a major disruption on economic activity.

We also harbored such grave concerns until we heard an American diplomat talk about the resale of burned-out light bulbs in the black (or just gray) markets of Moscow. Light bulbs were scarce in Moscow under the communist regime, partly because of the inefficiency of the Russian light-bulb producers and partly because light bulbs were underpriced and producers had only weak incentives to meet customer needs. To get light bulbs, Russians had to wait in long lines, possibly two or three hours. Reducing the shortage was impaired by the fact that many producers still did not have the right to own, buy, and sell all of the materials that go into light bulbs and the light bulbs themselves.

The planners, however, forgot about imposing such restrictions on the ownership and resale of used light bulbs, which were of no use to anyone, or so it might have been thought. Russian consumers found a use for them, however. They could buy the burned-out light bulbs, take them to work, and exchange them with good bulbs in their work places. They could then call the maintenance department to have the bulbs replaced. The bulbs are typically replaced with unusual quickness. Why? Because the maintenance people knew that they could claim ownership of the used bulbs, once they replaced them with good bulbs, and they could then sell the used bulbs back to the Moscow black market. The diplomat reported that used bulbs had a life cycle of approximately twenty-four hours. Within that time, the used bulbs would be back for resale again. To

---

[20] As described by Edwin Chadwich, "Opening Address," *Journal of the Royal Statistical Society of London*, vol. 25 (1862), as cited in Robert B. Ekelund and Robert F. Hebert, *A History of Economic Theory and Method*.

[21] Mikhail Gorbachev, *Perestroika* (New York: Harper & Row, 1987), p. 97.

paraphrase Gorbachev, amazing, unexpected things happen when people are given meaningful incentives to lay claims to the benefits of their actions.

Like most other universities, the University of California, Irvine, has graduate student apartments that are heavily subsidized. That is to say, the rents charged for on-campus apartments are several hundred dollars lower than the rents charged for comparable off-campus, privately owned apartments close to campus. The university claims that the apartments must be underpriced in order that "good" but "poor" graduate students can afford to follow their degree programs at the university. The university also argues that if it were to raise the rents, the quality of the graduate students the university could attract would fall unless graduate student assistantship and fellowship payments were raised.

Naturally, the incentives in the subsidized rents lead to consequences that undermine the official explanation for the subsidies. First, students will likely use the subsidies to rent apartments that are larger than they would choose to rent if they had to pay market prices, soaking up space that could be used by other students. Second, the quality of the apartments has deteriorated with deferred maintenance, which has been made necessary by the low rents. Third, and perhaps most important, the graduate students tend to stay much longer than you would think graduate students would need in order to finish their degree programs. Indeed, many of the student-residents have been in their apartments for more than a decade, using all sorts of means to prolong their graduations (for example, sending spouses to school, taking years off in the middle of their programs, and pursuing post-doctorate research). In extending their stays, they deny the spaces to other students who might otherwise choose UC-Irvine.

Moreover, we must question the official argument—they "can't afford higher rents"—for maintaining the low rents. Of course, the students could afford higher rents. If there is a problem, the university could raise the rent and hand the additional revenue back to the graduate students in the form of cash. If the rent were raised by $400 and the students were given the $400 back, then they could clearly afford what they had. The question is whether they would actually continue to rent the same apartments. Not likely. Many students would take the cash and run to buy other things, after accepting a smaller place to live (because the price of space would then be higher).

Instead of increasing the quality of the university's graduate students, the rent subsidies are very likely lowering the quality. If the rent were raised by $400 and the revenue were transferred to departments for distribution as assistantships and fellowships, then surely potential graduate students would be happier by having the $400 in cash than $400 in rent subsidy. With the cash, the students could still rent the apartment at the higher payment, but then they could do other more valuable things with the cash. When the subsidy is in the form of a reduction in the price of a particular good, then it is locked in, limited to the good in question, a point that has escaped the thinking of the university officials. This is one general reason why businesses—not just universities—should think seriously before they give their workers in-kind work-related benefits in lieu of salary.

Tying Pay to Performance

Of course, incentives have been found to be important for more mundane, everyday business reasons. Tying compensation to some objective measure of firm performance can cause the affected workers' productivity to rise substantially, a point that is covered in detail later. As would be expected, appropriately structured incentive pay can increase a firm's rate of return and stock price, as well as the income of the affected workers. One study of thousands of managers of large corporations found that adding a 10 percent bonus for good performance could be expected to add .3 to .9 percent to the companies' after-tax rate of return on stockholder investment. If the managerial bonuses are tied to the market prices of the companies' stock, share prices can be expected to rise by 4 to 12 percent. The study also found that the greater the sensitivity of management pay to company performance, the better the performance.[22] Another study found that firms don't have to wait around for the incentives to have an impact on the firms' bottom line to get a jump in their stock prices; all they have to do is *announce* that executives' compensation over the long haul is going to be more closely tied (through stock options or bonuses) to performance measures and the stock will, within days, go up several percentage points, increasing shareholder wealth by tens, if not hundreds, of millions of dollars (depending on firm size).[23]

Of course, if managers are paid just a straight salary, they have little reason to take on risky investments. They gain nothing from the higher rates of return associated with risky investments, which is why they may shy away from them. Accordingly, it should surprise no one to learn that when managers are given bonuses based on performance, they tend to undertake riskier, higher paying investments.[24] But then, if the bonuses are based on some short-term goal—say, this year's earnings—instead of some longer-term goal—say, some level for the stock price—you can bet that managers will tend to sacrifice investments with higher longer-term payoffs for smaller payoffs that are received within the performance period. The managers' time horizons can be lengthened by tying their compensation to the firm's stock value and then requiring that they hold the firm's stock until some later date, for example, retirement.[25]

Although incentives have always mattered, they probably have never been more important to businesses interested in competing aggressively on a global scale. Greater global competition means that producers everywhere must meet the best production

---

[22]The study covered the pay of 16,000 managers from 250 large corporations over the 1982-1986 period (John M. Abowd, "Does Performance -Based Managerial Compensation Affect Corporate Performance?" *Industrial and Labor Relations Review*, vol. 43 [special issue, February 1990], pp. 52S–3S).

[23]See James A. Brickley, Sanjai Bhagat, and Ronald C. Lease, "The Impact of Long-Range Managerial Compensation Plans on Shareholder Wealth," *Journal of Accounting and Economics*, vol. 7 (1985), pp. 115–29.

[24]Y. Amihud and B. Lev, "Risk Aversion as a Managerial Motive for Conglomerate Mergers," *Bell Journal of Economics* (Fall 1981), pp. 605–17; B. Holmstron, "Moral Hazard and Observability," *Bell Journal of Economics*, vol. 10 (1979), pp. 74–1; S. Shavell, "Risk Sharing and Incentives in the Principal and Agent Relationship," *Bell Journal of Economics*, vol. 10 (1979), pp. 55–3; and C. Smith and R. Watts, "Incentive and Tax Effects of Executive Compensation Plans," Australian Journal of Management, vol. 7 (1982), pp. 139–57.

[25]Michael C. Jensen and William H. Meckling, "Property Rights and Production Functions: An Application of Labor-Managed Firms and Codetermination," *Journal of Business*, vol. 52 (1979), pp. 469–06.

standards anywhere on the globe, which requires having the best incentive systems anywhere. Incentives will continue to grow in importance in business as the economy becomes more complex, more global, and more competitive. Although incentives are both positive and negative, when structured properly, incentives can ensure that managers, workers, and consumers prosper.

Like it or not, business people will have to learn to think about incentives with the same rigor that they now contemplate their balance sheets and marketing plans. They will need to justify the incentive structures they devise, which means they will have to understand why they do what they do. High pay and golden parachutes for executives and stock options for workers will need to be used judiciously. They can't be employed just because they seem like a nice idea, or because everyone else is using them. Investors who find it easier and easier to move their investment funds anywhere in the world will not allow their capital to be used for "nice ideas." Unless well thought out, "nice ideas" can spell wasted investments. The multitude of ways that incentives can matter in business must be incredibly large, which makes a study of them mandatory—if managers want to get them right.

Unless policies are carefully considered, perverse incentives can be an inadvertent consequence, mainly because people can be very creative in responding to policies. Lincoln Electric is known for achieving high productivity levels among its production workers by tying their pay to measures of how much they produce. But the company went too far. When it tied the pay of secretaries to "production," with counters installed on typewriters to measure how much was typed, the secretaries responded by spending their lunch hours typing useless pages of manuscript to increase their pay, which resulted in that incentive being quickly abandoned.[26] In seeking to reduce the number of "bugs" in its programs, a software company began paying programmers to find and fix bugs. The goal was noble but the response wasn't. Programmers began creating bugs in order that they could find and fix them, with one programmer increasing his pay $1,700 through essentially fraudulent means. The company eliminated the incentive pay scheme within a week of its introduction.[27] Incentives almost always work, but they don't always work well or in the way that's expected (a fact that has led to harsh criticisms of even attempting to use incentives, punishments, or rewards[28]).

In the twenty-first century world economy, business incentives will be commonplace; getting them right will be an abiding and taxing concern of managers.

The Role of Incentives in Firm Successes and Failures

Some firms prosper while other firms fail. Why? An easy answer is that some firms produce a better product or provide a better service. The fortunes of many fast-food

---

[26]See N. Fast and N. Berg, "The Lincoln Electric Company," Harvard Business School Case (Cambridge, Mass.: Harvard Business School Press, 1971).
[27]S. Adams, "Manager's Journal: The Dilbert Principle," *Wall Street Journal*, May 22, 1995, p. 14.
[28]For criticisms of incentives, see Alfie Kohn, *Punished by Rewards* (Boston: Houghton Mifflin, 1993). See also Jone L. Pearce, "Why Merit Pay Doesn't Work: Implications for Organization Theory," *Perspectives on Compensation* (1987).

restaurants have depended upon the quality of their burgers and the cleanliness of their rest rooms.

Some firms have failed not because they have done anything "wrong," but rather they have not done as much "right" as have their competitors. Many textile firms in the southeast part of the country have folded over the last two decades in spite of their substantial efforts to improve their productivity and increase quality. The failing firms closed their doors simply because they were not able meet the competition from lower-priced textile imports and from textiles produced by even more aggressive (and successful) domestic textile firms.[29]

Many firms have failed because they did not pay attention to their costs or because their managers were not very smart in setting their firms' product and service strategies to meet the changes in their markets. Even the 9/11 ctatrophy, several major airlines (and scores of smaller ones) have folded their wings over the last two decades because their planes and personnel were too expensive relative to the value of the service they provided and therefore relative to the prices they could charge in deregulated skies.

We agree that a lot of things are important to success in business, not the least of which are the leadership of managers, worker skills and character, firm strategies, and cost-control methods. One of the more important points managers must remember is that incentives can be very powerful forces within a firm—for good and bad! This means managers must pay attention to the art and logic of getting incentives *right*. In the "Manager's Corner" sections that are included in every chapter, we will examine a large number of different questions related to the organization of production within firms, most of which relate to incentives in one way or another: *How large should firms be? Do workers want tough bosses? Why don't more firms pay piece rate? What difference does debt make? What good are corporate raiders?* At the most obvious level, these questions are concerned with widely different problems firms have to face. But underneath all that is written about firm structure or piece-rate pay or corporate raiders in the "Manager's Corner" sections is an important theme: *Develop incentives so that everyone in your firm or connected to it—owners, executives, managers, workers, suppliers, and customers—win from your firm's operation.*

It is all too common for people to think that the only way for one group of "stakeholders" in a firm to gain is for some other group to lose. The search is all too frequently for ways to cut costs for one group of stakeholders (owners or managers) by skewering another group (line workers or customers). In this book we seek incentive arrangements by which everyone profits. That means that we seek incentives that are *mutually* beneficial, which necessarily means incentives that promote cooperation between everyone with an interest in the firm. Devising mutually beneficial incentives is a tough order, but we think it is the only way to ensure a viable business. Business arrangements that do not benefit all parties involved are arrangements that are not likely to survive for long.

---

[29]Indeed, many textile firms have failed because the expanding nontextile economies of their regions have pushed up labor costs, outcompeting some textile firms for the resources they need for continued production.

As noted, we typically think of firms competing with each other by producing better products at lower costs and making them more conveniently available to consumers at lower prices. But underlying this competition that we can observe in the marketplace is a more fundamental struggle taking place *within* firms to organize production in the most efficient manner, which necessarily requires an understanding of the incentives that face firm stakeholders—owners, managers, workers, and suppliers. The "Manager's Corners" have been written with one central proposition in mind: *In the competitive marketplace, the firms that survive and thrive are the ones that recognize that incentives matter—and they matter a great deal.* Successful firms play to the power of smart incentives (those that drive firm and worker incomes upward) and avoid perverse incentives (ones that undermine firm and worker incomes). And managers have good reason to make incentives a major focus for their firms: They can reduce their chances of being replaced.[30]

Why Incentives Are Important

But such facts beg a critical question, Why are incentives important? Why do they work? Admittedly, the answers are many. One of the more important reasons that incentives matter within firms is that firms are collections of workers whose interests are not always aligned with the interests of the people who employ them, that is, the owners. The principal problem facing the owners is how to get the workers to do what the owners want them to do. The owners could just issue directives, but without some incentive to obey the directives, nothing may happen. Directives may have some value in themselves; people do feel a sense of obligation to do what they were hired to do, and one of the things they may have been hired to do is obey orders (within limits). However, directives can be costly. Firms may use incentives simply as a cheaper substitute for giving out orders that can go unheeded unless the workers have some reason to heed them.

Firms may also use incentives to clarify firm goals, to spell out in concrete terms to workers what the owners want to accomplish. As every manager knows all too well, it's difficult to establish and write out the firm's strategy that will be used to achieve its stated goals, and it is an even more difficult task to get workers to appreciate, understand, and work toward those goals. The communication problem typically escalates with the size of the organization. Goals are always imperfectly communicated, especially by memoranda or through employment manuals that may be read once and tossed. Workers don't always know how serious the owners and upper managers are; they can remember any number of times when widely circulated memos were nothing but window dressing. Incentives are a means by which owners and upper managers can validate overall company goals and strategies. They can in effect say through incentives, "This is what we think is important. This is what we will be working toward. This is what we will be trying to get everyone else to do. And this is where we will put our money." Even if

[30] According to econometric research, those firms in the lowest decile of industry performance measured by profit and stock price increases were about 1.5 times as likely to have a change of top executives as firms in the best decile of profit and stock price performers. See M. Weisback, "Outside Directors and CEO Turnover," *Journal of Financial Economics*, vol. 20 (1988), pp. 431–60; and J. Warner, R. Watts, and K. Wruck, "Stock Prices and Top Management Changes," *Journal of Financial Economics*, vol. 20, pp. 461–92.

workers were not sensitive to the pecuniary benefits of work, but were only interested in doing what their companies wanted them to do, incentives, because of the messages they convey, can have a valued and direct impact on what workers do and how long and hard they work.[31]

But there is a far more fundamental reason that incentives matter: *Managers don't always know what orders or directives to give.* No matter how intelligent, hard working and well-informed managers are, they seldom know as much about particular jobs as those who are actually doing those jobs. Knowing about the peculiarities of a machine, the difficulties a fellow worker on the production line is experiencing at home, or the personality quirks of a customer are just a few examples of the innumerable particular bits of localized knowledge that are crucial to the success of a firm. And this knowledge is spread over everyone in the firm without the possibility of its being fully communicated to, and effectively utilized by, those who are primarily responsible for managerial oversight. The only way a firm can fully benefit from such localized knowledge is to allow those who possess the knowledge—the firm's employees—the freedom to use what they know.

Management theorists are increasingly recognizing this simple fact—that a great deal of knowledge is widely dispersed throughout the firm. In doing so, they are turning away from the approach to management recommended by Frederick Taylor.[32] At the beginning of the twentieth century, Taylor had popularized the time-and-motion approach to management in which experts, or managers, determined the most efficient way to do particular jobs and then required employees to work accordingly. Instead of the top-down or command style recommended by Taylor, the management profession is now sympathetic to a more participatory managerial approach, under which the management hierarchy is flatter, with authority for particular decisions dispersed throughout the firm, residing with those who are in the best position to exercise it. As noted, in varying degrees, all firms are necessarily involved in *participatory management* with practically everyone having some management authority over some firm resources. The principal difference between those workers at the top and bottom of the firm hierarchy is the scope of authority over resources.

But the benefits from participatory management can only be realized if employees have not only the freedom but also the motivation to use their special knowledge in productive cooperation with each other. The crucial ingredient for bringing about the requisite coordination is incentives that align the otherwise conflicting interests of individual employees with the collective interests of all members of the firm. Without such incentives, there can be no hope that the knowledge dispersed throughout the firm will be used in a cooperative and coordinated way. The only practical alternative to a

---

[31] This perspective on incentives is developed by Harrison C. White, "Agency as Control," *Principals and Agents: The Structure of Business*, edited by John W. Pratt and Richard J. Zeckhauser (Boston, Mass.: Harvard Business School Press, 1991), pp. 187–12; and James A. Robins, "Why and When Does Agency Theory Matter? A Critical Approach to the Role of Agency Theory in the Analysis of Organizational Control" (Irvine, Calif.: Graduate School of Management, University of California, Irvine, working paper, 1996).

[32] Frederick Taylor, *The Principles of Scientific Management* (New York: Harper, 1929).

functioning system of incentives is, again, a top-down, command-and-control approach that, unfortunately, can never allow the full potential of a firm's employees to be realized.

Managers must heed the words of social philosopher Friedrich Hayek, "The more men know, the smaller the share of knowledge becomes that any one mind [the planner's mind included] can absorb. The more civilized we become, the more relatively ignorant must each individual be of the facts on which the working of civilization depends. The very division of knowledge increases the necessary ignorance of the individual of most of this knowledge."[33] That insight applies within the firm. With the growing complexity and sophistication of production, knowledge becomes ever more widely dispersed among a growing number of workers. Hence, the importance of incentives has grown with modern-day leaps in the technological sophistication of products and production processes. Incentives will continue to grow in importance as production and distribution processes become ever more complex.

Seen in this light, the problem of the firm is the same as the problem of the general economy. As did Hayek, economists have argued for years that no group of government planners, no matter how intelligent and dedicated, can acquire all the localized knowledge necessary to allocate resources intelligently. The long and painful experiments with socialism and its extreme variant, communism, have confirmed that this is one argument that economists got right. But the freedom for people to use the knowledge that only they individually have has to be coupled with incentives that motivate people to use that knowledge in socially cooperative ways—meaning that the best way for individuals to pursue their own objectives is by making decisions that improve the opportunities for others to pursue their objectives. In a market economy these incentives are found primarily in the form of prices that emerge out of the rules of private property and voluntary exchange. Market prices provide the incentive people need to productively coordinate their decisions with each other, thus making it not only possible, but desirable, for people to have a large measure of freedom to make use of the localized information and know-how they have.

A perfect incentive system would assure that everyone could be given complete freedom because it would be in the interest of each to advance the interests of all. No such perfect incentive system exists, not within any firm or within any economy. In every economy there is always some appropriate mix of both market incentives and government controls that achieve the best overall results. The argument over just what the right mix is will no doubt continue indefinitely, but few deny that both incentives and controls are needed. Similarly, for any firm made up of more than one person, there is some mix of incentives and direct managerial control that best promotes the objectives of the firm; i.e., the general interests of its members.

Granted, incentives may not seem to matter much at any point in time, but even so, the power of incentives can accumulate with time. For example, suppose that without improved incentives firm profits will grow in real-dollar terms by 2 percent a year. Suppose that with more effective incentives firm profits can grow by 2.5 percent a year. The difference is not "much," just a half of a percentage point per year. However, the compound impact of the higher growth rate will mean that after 30 years, real profits will

---

[33] F. A. Hayek, *The Constitution of Liberty* (Chicago: University of Chicago Press, 1960), p. 26.

be 33 percent higher with the improved incentives (a fact that is likely to be reflected in current stock prices). Furthermore, the firm may be able to achieve the relatively higher profits with little or no cost. "Good" incentives may be no more expensive than "bad" incentives. Good incentives are the proverbial "free lunch" that economists typically dismiss.

Of course, if a given firm doesn't pay attention to its incentives, it may lose more than its lunch; it may be forced out of business by those firms that do recognize the importance of incentives. Seen from this perspective, incentives can be a critical component of firm survival, perhaps just as critical as product development or technological sophistication.

The problem is in getting the incentives right and using the full range of potential incentives. Unfortunately, we can't say exactly what incentives your firm should employ. The exact incentives chosen depend on local conditions that can vary greatly across firms. You would not want us to write about *particular incentives* for your particular circumstances, mainly because we can be assured of only one constant fact about business: Particular circumstances will change with time and markets. Here, we offer a *way of thinking* about incentives that, if employed with diligence, will enable managers and owners to get their firms' incentives more in line with their desire for increased productivity and profits.

Why Designated Hitters Get Hit

Admittedly, there is no way that managers can ever know for sure what the best set of incentives is. The problems of determining the proper incentives are many. And one of the main problems is not the dearth, but the great variety of incentives that can be used. Under the "Manager's Corner" sections, we necessarily focus on monetary incentives. This is mainly because such incentives have been well tested, but monetary incentives should be expected to be effective for the broad sweep of managers and workers: Most people can usually find some reason to want more money, given that it can be used to buy so many things that people want. But our emphasis on monetary incentives doesn't mean that money is all that matters to people at work, and managers should realize that simple fact. Managers need to know what counts. We *know* money should count for most people at work simply because money can be used to buy so many things that are valued by workers. But what attributes of work can count? That's not always an easy question to answer. Not recognizing the question, however, and not looking for answers can have incentive consequences that are not expected.

To see this point, we take a sports example that involves people at work, albeit baseball players. Starting in 1973, the American League allowed "designated hitters" to bat for pitchers (who are, generally, poor at batting). What would you expect to be the consequence of such a workplace change? Three economists have reasoned that given that American League pitchers would not come up to bat, we should expect that more batters would be hit by errant pitches in the American League than in the National League. This is because the American League pitchers would not have to fear being hit themselves in retaliation. Hence, American League pitchers could be expected to deliberately hit more batters or to take more chances of coming closer to batters than

would be the case in the National League. Using sophisticated statistical methods, the economists found what they expected: since 1973 (after adjusting for other relevant factors that might affect hit batters), 10 to 15 percent more batters in the American League have been hit than in the National League.[34]

We remind you that "hits" and "pats" on the back can be important ways of increasing firm profits. However, there is a problem in talking about "hits" and "pats," or any other nonmoney attribute of the work environment, that must be kept in mind: Are "hits" and "pats" *goods* (something workers want) or *bads* (something they don't want)? Clearly, most people might want to avoid being hit by a baseball going 90 miles an hour, but what about "hits" that come close to being "pats"? Some workers might consider a "pat" on the back as a valued form of encouragement, whereas others might consider them to be an unwanted form of patronization or sexual overture (depending on exactly how and where the "pat" is given).

As complicated as these issues are, we can't avoid them, and managers would not get the pay they do if all such problems of "what counts" in the workplace were easily and readily solved. Psychology will always be a part of management precisely because it helps identify workers' likes and dislikes. Economics will always be a part of management because it can guide managers in making money by instituting and adjusting on the margin the combination of money and nonmoney incentives set out for workers. You can bet that we, the authors, also can show how the workers' willingness to trade off money for other attributes of the work environment (for example, common courtesies and respect) can increase firm profits and, at the same time, enhance worker welfare. That means that an unheralded job of managers is to stay attuned to what their workers want and then try to figure out how much they are willing to pay for what they want.

Another problem in the management of incentives is that no set of incentives is ever perfect, nor could it be. But even if managers knew the best incentive structure and how best to implement it, a serious incentive problem would remain. *What incentive should managers have to find the* best *set of incentives?* That's a tough but interesting question. An understanding of the structure of firms requires that we recognize the need to subject managers, as well as other employees, to the proper incentives. The need to impose the proper set of incentives on managers is also necessary for understanding firms' financial structure. For example, the question of what combination of debt and equity instruments is best for financing a firm cannot be answered properly without a consideration of managerial incentives.

Concluding Comments

Economics is a discipline best described as the study of human interaction in the context of scarcity. It is the study of how, individually and collectively, people use their scarce resources to satisfy as many their wants as possible. The economic method is founded in a set of presuppositions about human behavior on which economists construct theoretical models.

---

[34] Brian L. Goff, William F. Shughart, and Robert D. Tollison, "Batter Up! Moral Hazard and the Effects of the Designated Hitter Rule on Hit Batsmen," *Economic Inquiry*, vol. 35 (July 1997), pp. 555–61.

A major purpose of this book is to describe the analytical tools economists use and in that way show how they study human behavior in general. However, we stress how this method of thinking can be used to understand the ways people act business. After all, MBA students want (or should want) to know how economic methods can help them become better managers. Throughout the book, we use these methods of thinking in our search for improved incentives within firms. Almost everyone understands that firms can turn substantial profits by building the proverbial "better mousetrap." We intend to stress how money can be made from careful thinking about business issues, including how people are rewarded for their work and investments.

Review Questions

1. In the prison camp described on pages 4-6, rations were distributed equally. Why did trade within and among bungalows result?
2. Recall the priest who traded the cigarettes for cheese, and cheese for cigarettes, so that he ended up with more cigarettes than he had initially. Did someone else in the camp lose by the priest's activities? How was the priest able to end up better off than when he began? What did his activities do to the price of cheese in the different bungalows?
3. Theories may be defective, but economists continue to use them. Why?
4. A microeconomics book designed for MBA students could include theories more complex than those in this book. What might be the tradeoffs in dealing with more complex theories?
5. Most MBA students study in "groups." If you are not in a study group, imaging yourself in one. What incentive problems do these groups have to overcome? How has your group sought to overcome the incentive problems?

READING: "I, Pencil"

Leonard E. Read[35]

I am a lead pencil—ordinary wooden pencil familiar to all boys and girls and adults who can read and write. (My official name is "Mongol 482." My many ingredients are assembled, fabricated and finished by Eberhard Faber Pencil Company, Wilkes-Barre, Pennsylvania.)

Writing is both my vocation and my avocation; that's all I do.

You may wonder why I should write a genealogy. Well, to begin with, my story is interesting. And, next, I am a mystery—more so than a tree or a sunset or even a flash of lightning. But, sadly, I am taken for granted by those who use me, as if I were a mere incident and without background. This supercilious attitude relegates me to the level of the commonplace. This is a species of the grievous error in which mankind cannot too long persist without peril. For, as a wise man, G.K. Chesterton, observed, "We are perishing for want or wonder, not for want of wonders."

I, Pencil, simple though I appear to be, merit your wonder and awe, a claim I shall attempt to prove. In fact, if you can understand me—no, that's too much to ask of anyone—if you can become aware of the miraculousness that I symbolize, you can help save the freedom mankind is so unhappily losing. I have a profound lesson to teach. And I can teach this lesson better than can an automobile or an airplane or a mechanical dishwasher because—well, because I am seemingly so simple.

Simple? Yet, not a single person on the face of this earth knows how to make me. This sounds fantastic, doesn't it? Especially when you realize that there are about one and one-half billion of my kind produced in the U.S. each year.

Pick me up and look me over. What do you see? Not much meets the eye—there's some wood, lacquer, the printed labeling, graphite lead, a bit of metal, and an eraser.

Innumerable Antecedents

Just as you cannot trace your family tree back very far, so is it impossible for me to name and explain all my antecedents. But I would like to suggest enough of them to impress upon you the richness and complexity of my background.

My family tree begins with what in fact is a tree, a cedar of straight grain that grows in Northern California and Oregon. Now contemplate all the saws and trucks and rope and the countless other gear used in harvesting and carting the cedar logs to the railroad siding. Think of all the persons and the numberless skills that went into their fabrication: the mining of ore, the making of steel and its refinement into saws, axes, motors; the growing of hemp and bringing it through all the stages to heavy and strong rope; the logging camps with their beds and mess halls, the cookery and the raising of all the foods. Why, untold thousands of persons had a hand in every cup of coffee the loggers drink!

The logs are shipped to a mill in San Leandro, California. Can you imagine the individuals who make flat cars and rails and railroad engines and who construct and install the communication systems incidental thereto? These legions are among my antecedents.

Consider the millwork in San Leandro. The cedar logs are cut into small, pencil-length slats less than one-fourth of an inch in thickness. These are kiln-dried and then tinted for the same reason women put rouge on their faces. People prefer that I look pretty, not a pallid white. The slats are waxed and kiln-dried again. How many skills went into the making of the tint and kilns, into supplying the heat, the light and power, the belts, motors, and all the other things a mill requires? Are sweepers in the mill among my ancestors? Yes, and also included are the men who poured the concrete for the dam of a Pacific Gas & Electric company hydroplant, which supplies the mill's power. And don't overlook the ancestors present and distant who have a hand in transporting sixty carloads of slats across the nation from California to Wilkes-Barre.

[3535] The late Mr. Reed was the founder of the Foundation for Economic Education. Permission for use in this volume granted by Donald Boudreaux, President, Foundation for Economic Education (May 4, 1999).

Complicated Machinery

Once in the pencil factory—$4,000,000 in machinery and building, all capital accumulated by thrifty and saving parents of mine—each slat is given eight grooves by a complex machine, after which another machine lays leads in every other slat, applies glue, and places another slat atop—a lead sandwich, so to speak. Seven brothers and I are mechanically carved from this "wood-clinched" sandwich.

My "lead" itself—it contains no lead at all—is complex. The graphite is mined in Ceylon. Consider the miners and those who make their many tools and the makers of the paper sacks in which the graphite is shipped and those who make the string that ties the sacks and those who put them aboard ships and those who make the ships. Even the lighthouse keepers along the way assisted in my birth—and the harbor pilots.

The graphite is mixed with clay from Mississippi in which ammonium hydroxide is used in the refining process. Then wetting agents are added such as sulfonated tallow—animal fats chemically reacted with sulfuric acid. After passing through numerous machines, the mixture finally appears as endless extrusions—as from a sausage grinder—cut to size, dried, and baked for several hours at 1,850 degrees Fahrenheit. To increase their strength and s moothness the leads are then treated with a hot mixture, which includes candililla wax from Mexico, paraffin wax and hydrogenated natural fats.

My cedar receives six coats of lacquer. Do you know all of the ingredients of lacquer? Who would think that the growers of castor beans and the refiners of castor oil are a part of it? They are. Why, even the processes by which the lacquer is made a beautiful yellow involves the skills of more persons than one can enumerate!

Observe the labeling. That's a film formed by applying heat to carbon black mixed with resins. How do you make resins and what, pray, is carbon black?

My bit of metal—the ferrule—is brass. Think of all the persons who mine zinc and copper and those who have the skills to make shiny sheet brass from these products of nature. Those black rings on my ferrule are black nickel. What is black nickel and how is it applied? The complete story of why the center of my ferrule has no black nickel on it would take pages to explain.

Then there's my crowning glory, inelegantly referred to in the trade as "the plug," the part man uses to erase the errors he makes with me. An ingredient called "factice" is what does the erasing. It is a rubber-like product made by reacting grape seed oil from the Dutch East Indies with sulfur chloride. Rubber, contrary to the common notion, is only for binding purposes. Then, too, there are numerous vulcanizing and accelerating agents. The pumice comes from Italy; and the pigment that gives "the plug" its color is cadmium sulfide.

Vast Web of Know-How

Does anyone wish to challenge my earlier assertion that no single person on the face of this earth knows how to make me?

Actually, millions of human beings have had a hand in my creation, no one of whom even knows more than a very few of the others. Now, you may say that I go too far in relating the picker of a coffee berry in far-off Brazil and food growers elsewhere to my creation; that this is an extreme position. I shall stand by my claim. There isn't a single person in all these millions, including the president of the pencil company, who contributes more than a tiny, infinitesimal bit of know-how. From the standpoint of know-how the only difference between the miner of graphite in Ceylon and the logger in Oregon is in the type of know-how. Neither the miner nor the logger can be dispensed with, any more than the chemist at the factory or the worker in the oil field—paraffin being a by-product of petroleum.

Here is an astounding fact: Neither the worker in the oil field nor the chemist nor the digger of graphite or clay nor anyone who mans or makes the ships or trains or trucks nor the one who runs the machine that does the knurling on my bit of metal nor the president of the company performs his singular task because he wants *me*. Each one wants me less, perhaps, than does a child in the first grade. Indeed, there are some

among this vast multitude who never saw a pencil nor would they know how to use one. Their motivation is other than me. Perhaps it is something like this: Each of these millions sees that he can thus exchange his tiny know-how for the goods and services he needs or wants. I may or may not be among these items.

No Human Master-Mind

There is a fact still more astounding: The absence of a master-mind, of anyone dictating or forcibly directing these countless actions that bring me into being. No trace of such a person can be found. Instead, we find the Scottish economist and moral philosopher Adam Smith's famous "Invisible Hand" at work in the marketplace. This is the mystery to which I earlier referred.

It has been said that "only God can make a tree." Why do we agree with this? Isn't it because we realize that we ourselves could not make one? Indeed, can we even describe a tree? We cannot, except in superficial terms. We can say, for instance, that a certain molecular configuration manifests itself as a tree. But what mind is there among men that could even record, let alone direct, the constant changes in molecules that transpire in the life span of a tree? Such a feat is utterly unthinkable!

I, Pencil, am a complex combination of miracles; a tree, zinc, copper, graphite, and so on. But to these miracles that manifest themselves in Nature an even more extraordinary miracle has been added: the configuration of creative human energies—millions of tiny bits of know-how configurating naturally and spontaneously in response to human necessity and desire and in the absence of any human master-minding! Since only God can make a tree, I insist that only God could make me. Man can no more direct millions of bits of know-how so as to bring a pencil into being than he can put molecules together to create a tree.

That's what I meant when I wrote earlier, "If you can become aware of the miraculousness that I symbolize, you can help save the freedom mankind is so unhappily losing." For, if one is aware that these bits of know-how will naturally, yes, automatically, arrange themselves into creative and productive patterns in response to human necessity and demand—that is, in the absence of governmental or any other coercive master-minding—then one will possess an absolutely essential ingredient for freedom: a faith in free men. Freedom is impossible without this faith.

Once government has had a monopoly on a creative activity—the delivery of the mail, for instance—most individuals will believe that the mail could not be efficiently delivered by men acting freely. And here is the reason: Each one acknowledges that he himself doesn't know how to do all the things involved in mail delivery. He also recognizes that no other individual could. These assumptions are correct. No individual possesses enough know-how to perform a nation's mail delivery any more than any individual possesses enough know-how to make a pencil. In the absence of a faith in free men—unaware that millions of tiny kinds of know-how would naturally and miraculously form and cooperate to satisfy this necessity—the individual cannot help but reach the erroneous conclusion that the mail can be delivered only by governmental master-minding.

Testimony Galore

If I, Pencil, were the only item that could offer testimony on what men can accomplish when free to try, then those with little faith would have a fair case. However, there is testimony galore; it's all about us on every hand. Mail delivery is exceedingly simple when compared, for instance, to the making of an automobile or a calculating machine or a grain combine or a milling machine, or to tens of thousands of other things.

Delivery? Why, in this age where men have been left free to try, they deliver the human voice around the world in less than one second; they deliver an event visually and in motion to any person's home when it is happening; they deliver 150 passengers from Seattle to Baltimore in less than four hours; they deliver gas from Texas to one's range or furnace in New York at unbelievably low rates and without subsidy; they deliver each four pounds of oil from the Persian Gulf to our Eastern Seaboard—halfway around the world—for less money than the government charges for delivering a one-ounce letter across the street!

Leave Men Free

The lesson I have to teach is this: Leave all creative energies uninhibited. Merely organize society to act in harmony with this lesson. Let society's legal apparatus remove all obstacles the best it can. Permit creative know-how to freely flow. Have faith that free men will respond to the "Invisible Hand." This faith will be confirmed. I, Pencil, seemingly simple though I am, offer the miracle of my creation as testimony that this is a practical faith, as practical as the sun, the rain, a cedar tree, and the good earth.

CHAPTER 2

Competitive Product Markets And Firm Decisions

Competition, if not prevented, tends to bring about a state of affairs in which: first, everything will be produced which somebody knows how to produce and which he can sell profitably at a price at which buyers will prefer it to the available alternatives: second, everything that is produced is produced by persons who can do so at least as cheaply as anybody else who in fact is not producing it: and third, that everything will be sold at prices lower than, or at least as low as, those at which it could be sold by anybody who in fact does not do so.

Friedrich Hayek

In the heart of New York City, Fred Lieberman's small grocery is dwarfed by the tall buildings that surround it. Yet it is remarkable for what it accomplishes. Lieberman's carries thousands of items, most of which are not produced locally, and some of which come thousands of miles from other parts of this country or abroad. A man of modest means, with little knowledge of production processes, Fred Lieberman has nevertheless been able to stock his store with many if not most of the foods and toiletries his customers need and want. Occasionally Lieberman's runs out of certain items, but most of the time the stock is ample. Its supply is so dependable that customers tend to take it for granted, forgetting that Lieberman's is one small strand in an extremely complex economic network.

How does Fred Lieberman get the goods he sells, and how does he know which ones to sell and at what price? The simplest answer is that the goods he offers and the prices at which they sell are determined through the market process- the interaction of many buyers and sellers trading what they have (their labor or other resources) for what they want. Lieberman stocks his store by appealing to the private interests of suppliers -- by paying them competitive prices. His customers pay him extra for the convenience of purchasing goods in their neighborhood grocery -- in the process appealing to his private interests. To determine what he should buy, Fred Lieberman considers his suppliers prices. To determine what and how much they should buy, his customers consider the prices he charges. The Nobel Prize-winning economist Friedrich Hayek has suggested that the market process is manageable for people like Fred Lieberman precisely because prices condense into usable form a great deal of information, signaling quickly what people want, what goods cost, and what resources are readily available. Prices guide and coordinate the sellers' production decisions and consumers' purchases.

How are prices determined? That is an important question for people in business simply because an understanding of how prices are determined can help business people understand

the forces that will cause prices to change in the future and, therefore, the forces that affect their businesses' bottom lines. There's money to be made in being able to understand the dynamics of prices. Our most general answer in this chapter to the question is deceptively simple: In competitive markets, the forces of supply and demand establish prices. However, there is much to be learned through the concepts of supply and demand. Indeed, we suspect that most MBA students will find supply and demand the most useful concepts developed in this book. However, to understand supply and demand, you must first understand the market process that is inherently competitive.

The Competitive Market Process

So far, our discussion of markets and their consequences has been rather casual. In this section we will define precisely such terms as market and competition. In later sections we will examine the way markets work and learn why, in a limited sense, markets can be considered efficient systems for determining what and how much to produce.

The Market Setting

Most people tend to think of a market as a geographical location -- a shopping center, an auction bar, a business district. From an economic perspective, however, it is more useful to think of a market as a process. You may recall from Chapter 1 that a market is defined as the process by which buyers and sellers determine what they are willing to buy and sell and on what terms. That is, it is the process by which buyers and sellers decide the prices and quantities of goods to be bought and sold.

In this process, individual market participants search for information relevant to their own interests. Buyers ask about the models, sizes, colors, and quantities available and the prices they must pay for them. Sellers inquire about the types of goods and services buyers want and the prices they are willing to pay.

This market process is self-correcting. Buyers and sellers routinely revise their plans on the basis of experience. As Israel Kirzner has written,

> The overly ambitious plans of one period will be replaced by more realistic ones; market opportunities overlooked in one period will be exploited in the next. In other words, even without changes in the basic data of the market, the decision made in one period onetime generates systematic alterations in corresponding decisions for the succeeding period.[1]

The market is made up of people, consumers and entrepreneurs, attempting to buy and sell on the best terms possible. Through the groping process of give and take, they move from relative ignorance about others' wants and needs to a reasonably accurate understanding of

[1] Israel Kirzner, Competition and Entrepreneurship (Chicago: University of Chicago Press, 1973), p. 10.

how much can be bought and sold and at what price. The market functions as an ongoing information and exchange system.

Competition Among Buyers and Among Sellers

Part and parcel of the market process is the concept of competition. **Competition** is the process by which market participants, in pursuing their own interests, attempt to outdo, outprice, outproduce, and outmaneuver each other. By extension, competition is also the process by which market participants attempt to avoid being outdone, outpriced, outproduced, or outmaneuvered by others.

Competition does not occur between buyer and seller, but among buyers or among sellers. Buyers compete with other buyers for the limited number of goods on the market. To compete, they must discover what other buyers are bidding and offer the seller better terms -- a higher price or the same price for a lower-quality product. Sellers compete with other sellers for the consumer's dollar. They must learn what their rivals are doing and attempt to do it better or differently -- to lower the price or enhance the product's appeal.

This kind of competition stimulates the exchange of information, forcing competitors to reveal their plans to prospective buyers or sellers. The exchange of information can be seen clearly at auctions. Before the bidding begins, buyer look over the merchandise and the other buyers, attempting to determine how high others might be willing to bid for a particular piece. During the auction, this specific information is revealed as buyers call out their bids and others try to top them. Information exchange is less apparent in department stores, where competition is often restricted. Even there, however, comparison-shopping will often reveal some sellers who are offering lower prices in an attempt to attract consumers.

In competing with each other, sellers reveal information that is ultimately of use to buyers. Buyers likewise inform sellers. From the consumer's point of view,

> The function of competition is here precisely to teach us who will serve us well: which grocer or travel agent, which department store or hotel, which doctor or solicitor, we can expect to provide the most satisfactory solution for whatever particular personal problem we may have to face.[2]

From the seller's point of view -- say, the auctioneer's -- competition among buyers brings the highest prices possible.

Competition among sellers takes many forms, including the price, quality, weight, volume, color, texture, poor durability, and smell of products, as well as the credit terms offered to buyers. Sellers also compete for consumers' attention by appealing to their hunger and sex drives or their fear of death, pain, and loud noises. All these forms of competition can be divided into two basic categories -- price and nonprice competition. Price competition is of particular interest to economists, who see it as an important source of information for market

---

[2] Friedrich H. Hayek, "The Meaning of Competition," Individualism *and Economic Order* (Chicago: University of Chicago Press, 1948), p. 97.

participants and a coordinating force that brings the quantity produced into line with the quantity consumers are willing and able to buy. In the following sections, we will construct a model of the competitive market and use it to explore the process of price competition. Nonprice competition will be covered in a later section.

Supply and Demand: A Market Model

A fully competitive market is made up of many buyers and sellers searching for opportunities or ready to enter the market when opportunities arise. To be described as competitive, therefore, a market must include a significant number of actual or potential competitors. A fully competitive market offers freedom of entry: there are no legal or economic barriers to producing and selling goods in the market.

Our market model assumes perfect competition-an ideal situation that is seldom, if ever, achieved in real life but that will simplify our calculations. **Perfect competition** is a market composed of numerous independent sellers and buyers of an identical product, such that no one individual seller or buyer has the ability to affect the market price by changing the production level. Entry into and exit from a perfectly competitive market is unrestricted. Producers can start up or shut down production at will. Anyone can enter the market, duplicate the good, and compete for consumers' dollars. Since each competitor produces only a small share of the total output, the individual competitor cannot significantly influence the degree of competition or the market price by entering or leaving the market.

This kind of market is well suited to graphic analysis. Our discussion will concentrate on how buyers and sellers interact to determine the price of tomatoes, a product Mr. Lieberman almost always carries. It will employ two curves. The first represents buyers' behavior, which is called their demand for the product.

The Elements of Demand

To the general public, demand is simply what people want, but to economists, demand has much more technical meaning. **Demand** is the assumed inverse relationship between the price of a good or service and the quantity consumers are willing and able to buy during a given period, all other things held constant.

Demand as a Relationship

The relationship between price and quantity is normally assumed to be inverse. That is, when the price of a good rises, the quantity sold, *ceteris paribus* (Latin for "everything else held constant"), will go down. Conversely, when the price of a good falls, the quantity sold goes up. Demand is not a quantity but a relationship. A given quantity sold at a particular price is properly called *quantity demanded.*

Both tables and graphs can be used to describe the assumed inverse relationship between price and quantity.

Demand as a Table or a Graph

Demand may be thought of as a schedule of the various quantities of a particular good consumers will buy at various prices. As the price goes down, the quantity purchased goes up and vice versa. Table 2.1 contains a hypothetical schedule of the demand for tomatoes in the New York area during a typical week. The middle column shows prices that might be charged. The column on the right shows the number of bushels consumers will buy at those prices. Note that as the price rises from zero to $11 a bushel, the number of bushels purchased drops from 110,000 to zero.

Demand may also be thought of as a curve. If price is scaled on a graph's vertical axis and quantity on the horizontal axis, the demand curve has a negative slope (downward and to the right), reflecting the assumed inverse relationship between price and quantity. The shape of the market demand curve is shown in Figure 2.1, which is based on the data from Table 2.1. Points *a* through *l* on the graph correspond to the price-quantity combinations *A* through *L* in the table. Note that as the price falls from P_2 ($8) to P_1 ($5), consumers move down their demand curve from a quantity of Q_1 (30) to the larger quantity Q_2 (60).[3]

The Slope and Determinants of Demand

Price and quantity are assumed to be inversely related for two reasons. First, as the price of a good decreases (and the prices of all other goods stay the same -- remember *ceteris paribus*), the purchasing power of consumer incomes rises. More consumers are able to buy the good, and many will buy more of most goods. (This response is called the income effect.)

In addition, as the price of a good decreases (and the prices of all other goods remain the same), the good becomes relatively cheaper, and consumers will substitute that good for others. (This response is called the substitution effect.)

---

[3] Mathematically, the demand relationship may be stated as $Q_d = a - bP$, where Q_d is the quantity demanded at every price; a is the quantity consumers will buy when the price is zero; b is the slope of the demand curve; and P is the price of the good. Thus the demand function for tomatoes described in Table 2.1 and Figure 2.1 may be written as $Q_d = 110{,}000 - 10{,}000\,P$.

TABLE 2.1 Market Demand for Tomatoes

Price-Quantity Combinations	Price per Bushel	Number of Bushels
A	$0	110,000
B	1	100,000
C	2	90,000
D	3	80,000
E	4	70,000
F	5	60,000
G	6	50,000
H	7	40,000
I	8	30,000
J	9	20,000
K	10	10,000
L	11	0

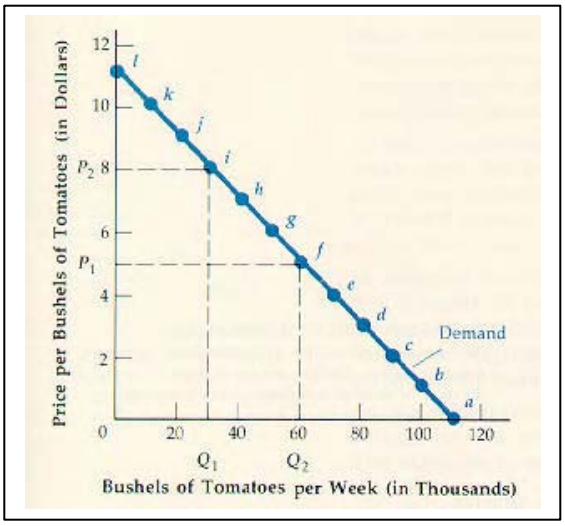


FIGURE 2.1 Market Demand for Tomatoes

Demand, the assumed inverse relationship between price and quantity purchased, can be represented by a curve that slopes down toward the right. Here, as the price falls from $11 to zero, the number of bushels of tomatoes purchased per week rises from zero to 110,000.

In sum, when the price of tomatoes (or razorblades or any other good) falls, more tomatoes will be purchased because more people will be buying them for more purposes.

Although price is an important part of the definition of demand, it is not the only determinant of how much of a good people will want. It may not even be the most important. The major factors that affect market demand are called determinants of demand. They are:

- Consumer tastes or preferences
- The prices of other goods
- Consumer incomes
- Number of consumers

- Expectations concerning future prices and incomes

A host of other factors, like weather, may also influence the demand for particular goods-ice cream, for instance.

A change in any of these determinants of demand will cause either an increase or a decrease in demand.

- An **increase in demand** is an increase in the quantity demanded at each and every price. It is represented graphically by a rightward, or outward, shift in the demand cure.
- A **decrease in demand** is a decrease in the quantity demanded at each and every price. It is represented graphically by a leftward, or inward, shift of the demand curve.

Figure 2.2 illustrates the shifts in the demand curve that result from a change in one of the determinants of demand. The outward shift from D_1 to D_2 indicates an increase in demand: consumers now want more of a good at each and every price. For example, they want Q_3 instead of Q_2 tomatoes at price P_2. Consumers are also willing to pay a higher price now for any quantity. For example, they will pay P_3 instead of P_2 for Q_2 tomatoes. The inward shift from D_1 to D_3 indicates a decrease in demand: consumers want less of a good at each and every price -- Q_1 instead of Q_2 tomatoes at price P_2. And they are willing to pay less than before for any quantity -- P_1 instead of P_2 for Q_2 tomatoes.

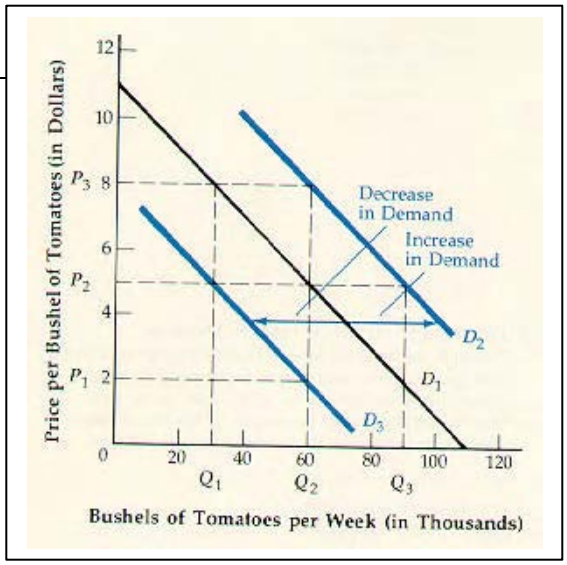


FIGURE 2.2 Shifts in the Demand Curve

An increase in demand is represented by a rightward, outward, shift in the demand curve, from D_1 to D_2. A decrease in demand is represented by a leftward, or inward, shift in the demand curve, from D_1 to D_3.

A change in a determinant of demand may be translated into an increase or decrease in market demand in numerous ways. An increase in market demand can be caused by:

An increase in consumers' desire for the good. If people truly want the good more, they will buy more of the good at any given price or pay a higher price for any given quantity.

An increase in the number of buyers. If people will buy more of the good at any given price, they will also pay a higher price for any given quantity.

An increase in the price of substitute goods (which can be used in place of the good in question). If the price of oranges increases, the demand for grapefruit will increase.

A decrease in the price of complement goods (which are used in conjunction with the good in question). If the price of stereo systems falls, the demand for records, tapes, and CDs will rise.

Generally speaking (but not always), an increase in consumer incomes. An increase in people's incomes may increase the demand for luxury goods, such as new cars. It may also decrease demand for low-quality goods (like hamburger) because people can now afford better-quality products (like steak).

An expected increase in the future price of the good in question. If people expect the price of cars to rise faster than the prices of other goods, then (depending on exactly when they expect the increase) they may buy more cars now, thus avoiding the expected additional cost in the future.

An expected increase in the future price of a substitute good. If people expect the price of oranges to fall in the future, then (depending on exactly when they expect the price decrease) they may reduce their current demand for grapefruit, so they can buy more oranges in the future.

An expected increase in future incomes of buyers. College seniors' demand for cars tends to increase as graduation approaches and they anticipate a rise in income. The determinants of a decrease in market demand are just the opposite:

A decrease in consumers' desire or taste for the good.

A decrease in the number of buyers.

A decrease in the price of substitute goods.

An increase in the price of complement goods.

Usually (but not always), a decrease in consumer incomes.

An expected decrease in the future price of the good in question.

An expected decrease in the future price of a substitute good.

An expected decrease in the future incomes of buyers.

The Elements of Supply

On the other side of the market are producers of goods. The average person thinks of supply as the quantity of a good producers are willing to sell. To economists, however, supply means something quite different. **Supply** is the assumed relationship between the quantity of a good producers are willing to offer during a given period and the price, everything else held constant. Generally, because additional costs tend to rise with expanded production, this relationship is presumed to be positive. Like demand, supply is not a given quantity—that is called quantity supplied. Rather it is a relationship between price and quantity. As the price of a good rises, producers are generally willing to offer a larger quantity. The reverse is equally true: as price decreases, so does quantity supplied. Like demand, supply can also be described in a table or a graph.

Supply as a Table or a Graph

Supply may be described as a schedule of the quantity producers will offer at various prices during a given period of time. Table 2.2 shows such a supply schedule. As the price of tomatoes goes up from zero to $11 a bushel, the quantity offered rises from zero of 110,000, reflecting the assumed positive relationship between price and quantity.

Supply may also be thought of as a curve. If the quantity producers will offer is scaled on the horizontal axis of a graph and the price of the good is scaled on the vertical axis, the supply curve will slope upward to the right, reflecting the assumed positive relationship between price and quantity. In Figure 2.3, which was plotted from the data in Table 2.2, points *a* through *l* represent the price-quantity combinations A through L. Note how a change in the price causes a movement along the supply curve.[4]

The Slope and Determinants of Supply

The quantity producers will offer on the market depends on their production costs. Obviously the total cost of production will rise when more is produced because more resources will be required to expand output. The additional or marginal cost of each additional bushel produced also tends to rise as total output expands. In other words, it costs more to produce the second bushel of tomatoes than the first, and more to produce the third than the second. Firms will not expand their output unless they can cover their higher unit costs with a higher price. This is the reason the supply curve is thought to slope upward.

Anything that affects production costs will influence supply and the position of the supply curve. Such factors, which are called determinants of supply, include:

- Change in productivity due to a change in technology

[4] Mathematically, the supply relationship may be stated as $Q_s = a + bP$. Where Q_s is the quantity supplied; a is the quantity producers will supply when the price is zero; b is the slope; and P is the price. Thus the supply function of tomatoes represented in Table 2.2 and Figure 2.3 may be written $Q_s = 0 + 10{,}000$ P.

- Change in the profitability of producing other goods
- Change in the scarcity (and prices) of various productive resources

Many other factors, such as weather, can also affect production costs. A change in any of these determinants of supply can either increase or decrease supply.

- An **increase in supply** is an increase in the quantity producers are willing and able to offer at each and every price. It is represented graphically by a rightward, or outward, shift in the supply curve.
- A **decrease in supply** is a decrease in the quantity producers are willing and able to offer at each and every price. It is represented graphically by a leftward, or inward, shift in the supply curve.

TABLE 2.2 Market Supply of Tomatoes

Price-Quantity Combinations	Price per Bushel	Number of Bushels
A	$0	0
B	1	10
C	2	20
D	3	30
E	4	40
F	5	50
G	6	60
H	7	70
I	8	80
J	9	90
K	10	100
L	11	110

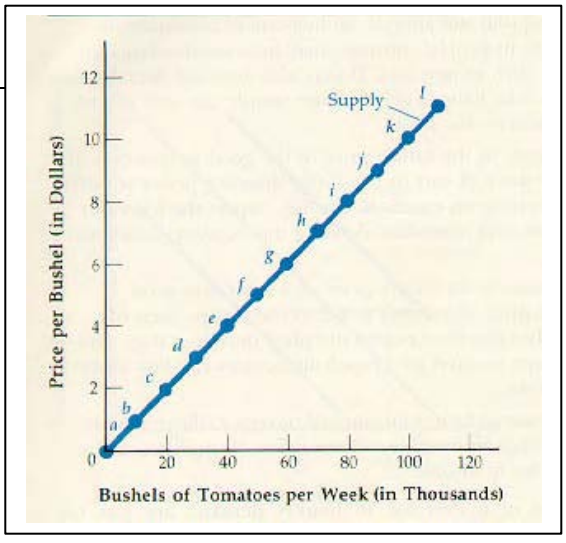


FIGURE 2.3 Supply of Tomatoes

Supply, the assumed relationship between price and quantity produced, can be represented by a curve that slopes up toward the right. Here, as the price rises from zero to $11, the number of bushels of tomatoes offered for sale during the course of a week rises from zero to 110,000.

In Figure 2.4, an increase in supply is represented by the shift from $S_1$to S_2. Producers are willing to produce a larger quantity at each price -- Q_3 instead of Q2 at price P_2, for example. They will also accept a lower price for each quantity -- P_1 instead of P2 for quantity Q_2. Conversely, the decrease in supply represented by the shift from S_1 to S_3 means that producers will offer less at each price -- Q_1 instead of Q_2 at price P_2. They must also have a higher price for each quantity -- P_3 instead of P_2 for quantity Q_2.

A few examples will illustrate the impact of changes in the determinants of supply. If firms learn how to produce more goods with the same or fewer resources, the cost of producing any given quantity will fall. Because of the technological improvement, firms will be able to offer a larger quantity at any given price or the same quantity at a lower price. The supply will increase, shifting the supply curve outward to the right.

Similarly, if the profitability of producing oranges increases relative to grapefruit, grapefruit producers will shift their resources to oranges. The supply of oranges will increase, shifting the supply curve to the right. Finally, if lumber (or labor or equipment) becomes scarcer, its price will rise, increasing the cost of new housing and reducing the supply. The supply curve will shift inward to the left.

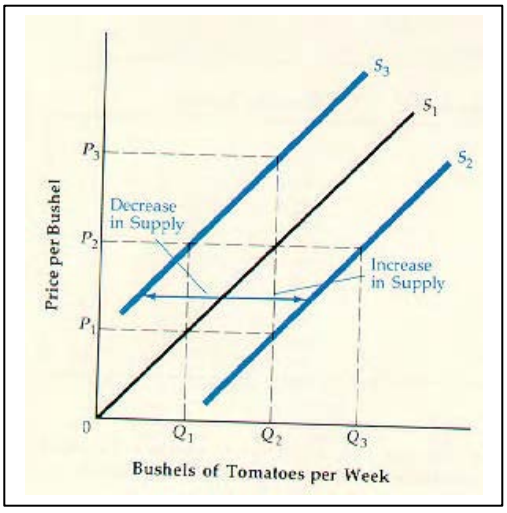


FIGURE 2.4 Shifts in the Supply Curve

A rightward, or outward, shift in the supply curve, from S_1 to S_2, represents an increase in supply. A leftward, or inward, shift in the supply curve, from S_1 to S_3, represents a decrease in supply.

Market Equilibrium

Supply and demand represent the two sides of the market—sellers and buyers. By plotting the supply and demand curves together, as in Figure 2.5 we can predict how buyers and sellers will be inconsistent, and a market surplus or shortage of tomatoes will result.

Market Surpluses

Suppose that the price of a bushel of tomatoes is $9, or P_2 in Figure 2.5. At this price the quantity demanded by consumers is 20,000 bushels, much less than the quantity offered by

producers, 90,000. There is a market surplus, or excess supply, of 70,000 bushels. A **market surplus** is the amount by which the quantity supplied exceeds the quantity demanded at any given price. Graphically, an excess quantity supplied occurs at any price above the intersection of the supply and demand curves.

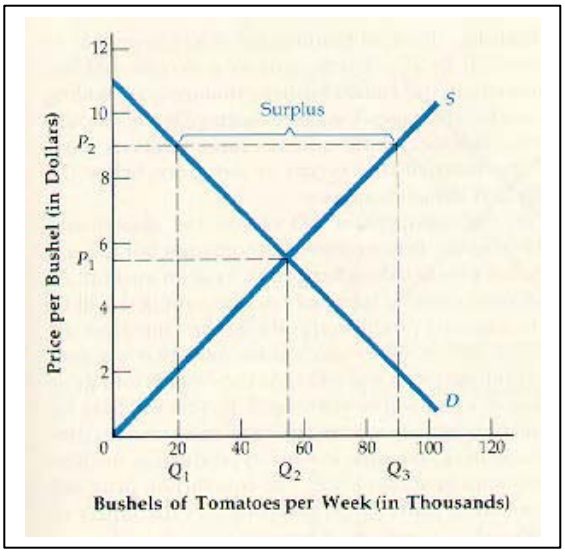


FIGURE 2.5 Market Surplus

If a price is higher than the intersection of the supply and demand curves, a market surplus—a greater quantity supplied, Q_3, than demanded, Q_1—results. Competitive pressure will push the price down to the equilibrium price P_1, the price at which the quantity supplied equals the quantity demanded (Q_2).

What will happen in this situation? Producers who cannot sell their tomatoes will have to compete by offering to sell at a lower price, forcing other producers to follow suit. As the competitive process forces the price down, the quantity consumers are willing to buy will expand, while the quantity producers are willing to sell will decrease. The result will be a contraction of the surplus, until it is finally eliminated at a price of $5.50 or P_1 (at the intersection of the two curves). At that price, producers will be selling all they want to; they will see no reason to lower prices further. Similarly, consumers will see no reason to pay more; they will be buying all they want. This point, where the wants of buyers and sellers intersect, is called the equilibrium price.

- The **equilibrium price** is the price toward which a competitive market will move, and at which it will remain once there, everything else held constant. It is the price at which the market "clears"—that is, at which the quantity demanded by consumers is matched exactly by the quantity offered by producers. At the equilibrium price, the quantities desired by buyers and sellers are also equal. This is the equilibrium quantity.
- The **equilibrium quantity** is the output (or sales) level toward which the market will move, and at which it will remain once there, everything else held constant.

In sum, a surplus emerges when the price asked is above the equilibrium price. It will be eliminated, through competition among sellers, when the price drops to the equilibrium price.

Market Shortages

Suppose the price asked is below the equilibrium price, as in Figure 2.6. At the relatively low price of $1, or P_1, buyers want to purchase 100,000 bushels—substantially more than the 10,000 bushels producers are willing to offer. The result is a market shortage. A **market shortage** is the amount by which the quantity demanded exceeds the quantity supplied at any given price. Graphically, it is the shortfall that occurs any price below the intersection of the supply and demand curves.

As with a market surplus, competition will correct the discrepancy between buyers' and sellers' plans. Buyers who want tomatoes but are unable to get them at a price of $1 will bid higher prices, as at an auction. As the price rises, a larger quantity will be supplied because suppliers will be better able to cover their increasing production costs. At the same time the quantity demanded will contract as buyers seek substitutes that are now relatively less expensive compared with tomatoes. At the equilibrium price of $5.50, or P_2, the market shortage will be eliminated. Buyers will have no reason to bid prices up further, for they will be getting all the tomatoes the want at that price. Sellers will have no reason to expand production further; they will be selling all they want to at that price. The equilibrium price will remain the same until some force shifts the position of either the supply or the demand curve. If such a shift occurs, the price will moves toward a new equilibrium at the new intersection of the supply and demand curves.

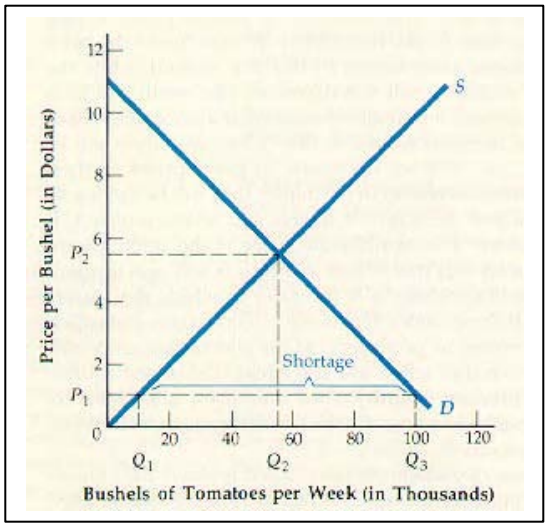


FIGURE 2.6 Market Shortages

A price that is below the intersection of the supply and demand curves will create a shortage—a greater quantity demanded, Q_3 than supplied Q_1. Competitive pressure will push the price up to the equilibrium price P_2, the price at which the quantity supplied equals the quantity demanded.

The Effect of Changes in Demand and Supply

Figure 2.7 shows the effects of shifts in demand and supply on the equilibrium price and quantity. In panel (a), an increase in demand from D_1 to D_2 raises the equilibrium price from $P_1$to P_2 and quantity from Q_2 to Q_1. Panel (b) shows the reverse effects of a decrease in demand.

An increase in supply from S_1 to S_2 -- panel (c) has a different effect. The equilibrium quantity rises from Q_1 to Q_2, but the equilibrium price falls from P_2 to P_1. A decrease in supply from $S_1$to S_2 -- panel (d) -- causes the opposite effect: the equilibrium quantity falls from Q_2 to Q_1, and the equilibrium price rises from P_1 to P_2.

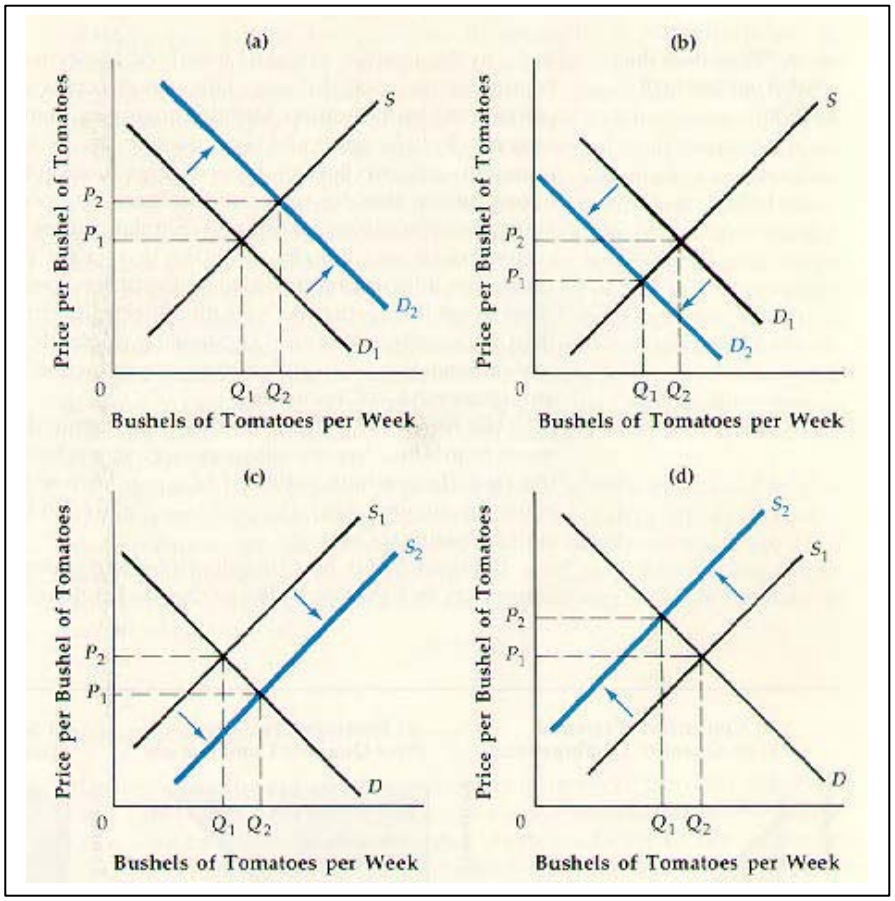


FIGURE 2.7 The Effects of Changes in Supply and Demand

An increase in demand—panel (a) -- raises both the equilibrium price and the equilibrium quantity. A decrease in demand -- panel (b) -- has the opposite effect: a decrease in the equilibrium price and quantity. An increase in supply -- panel (c)—causes the equilibrium quantity to rise but the equilibrium price to fall. A decrease in supply -- panel (d) -- has the opposite effect: a rise in the equilibrium price and a fall in the equilibrium quantity.

Price Ceilings and Price Floors

Political leaders have occasionally objected to the prices charged in open, competitive markets and have mandated the prices at which goods must be sold. That is, the government has enforced price ceilings and price floors. A **price ceiling** is a government-determined price above which a specified good cannot be sold. A **price floor** is a government-determined price below which a specified good cannot be sold. Supply and demand graphs can illustrate the consequences of price ceilings and floors. For example, some cities impose ceilings on the rents (or prices) for apartments. Such a ceiling must be below the equilibrium price—somewhere below P_1 in Figure 2.8(a). (If the ceiling were above equilibrium, it would be above the market price and would serve no purpose.) As the graph shows, such a price control creates a market shortage. The number of people wanting apartments, Q_2, is greater than the number of apartments available, Q_1. Because of the shortage, landlords will be less concerned about maintaining their units, for they will be able to rent them in any case.

If the government imposes a price *floor* -- on a commodity like milk, for example—the price must be above the equilibrium price, P_1 in Figure 2.8b. (A price floor below P_1 would be irrelevant, because the market would clear at a higher level on its own.) The result of such a price edict is a market surplus. Producers want to sell more milk, Q_2, than consumers are willing to buy, Q_1. Some producers -- those caught holding the surplus ($Q_2 - Q_1$) -- will be unable to sell all they want to sell. Eventually someone must bear the cost of destroying or storing the surplus -- and in fact the government holds vast quantities of its past efforts to support an equilibrium price for those products.

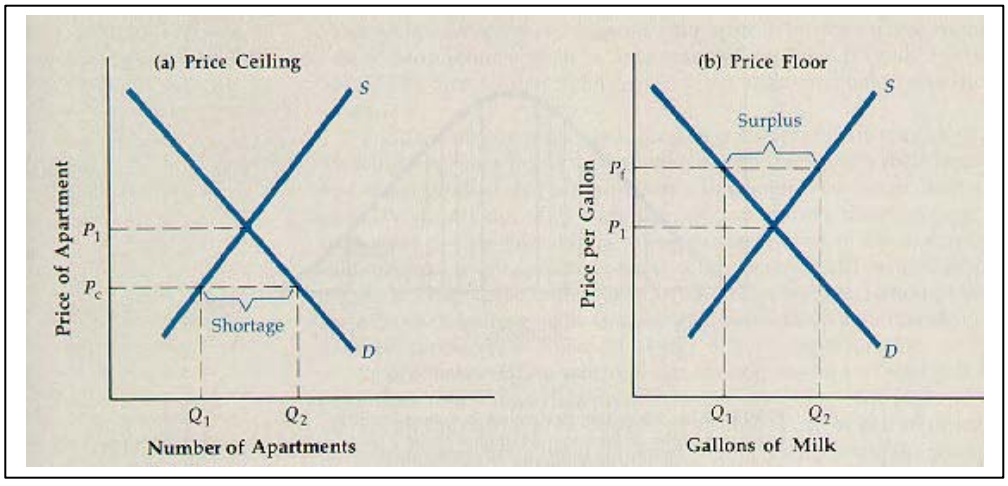


FIGURE 2.8 Price Ceilings and Floors

A price ceiling P_c—panel (a)—will create a market shortage equal to $Q_2 - Q_1$. A price floor P_f -- panel (b) -- will create a market surplus equal to $Q_2 - Q_1$.

The Efficiency of the Competitive Market Model

Early in this chapter we asked how Fred Lieberman knows what prices to charge for the goods he sells. The answer is now apparent: he adjusts his prices until his customers buy the quantities that he wants to sell. If he cannot sell all the fruits and vegetables he has, he lowers his price to attract customers and cuts back on his orders for those goods. If he runs short, he knows he can raise his prices and increase his orders. His customers then adjust their purchases accordingly. Similar actions by other producers and customers all over the city move the market for produce toward equilibrium. The information provided by the orders, reorders, and cancellations from stores like Lieberman's eventually reaches the suppliers of goods and then the suppliers of resources. Similarly wholesale prices give Fred Lieberman information on suppliers' costs of production and the relative scarcity and productivity of resources.

The use of the competitive market system to determine what and how much to produce has two advantages. First, it is tolerably accurate. Much of the time the amount produced in a competitive market system tends to equal the amount consumers want—no more, no less. Second, the market system maximizes output.

In Figure 2.9(a), note that all price-quantity combinations acceptable to consumers lie either on or below the market demand curve, in the shaded area. (If consumers are willing to pay P_2 for Q_1 then they should also be willing to pay less for that quantity—for example, P_1.) Furthermore, all price-quantity combinations acceptable to producers lie either on or above the supply curve, in the shaded area shown in Figure 2.9(b). (If producers are willing to accept P_1 for quantity Q_1, then they should also be willing to accept a higher price—for example, P_2). When supply and demand curves are combined in Figure 2.9(c), we see that all price-quantity combinations acceptable to both consumers and producers lie in the darkest shaded triangular area. From all those acceptable output levels, the competitive market produces Q_1, the maximum output level that can be produced given what producers and consumers are willing and able to do. In this respect, the competitive market can be said to be efficient, or to allocate resources efficiently. **Efficiency** is the maximization of output through careful allocation of resources, given the constraints of supply (producers' costs) and demand (consumers' preferences). The achievement of efficiency means that consumers' or producers' welfare will be reduced by an expansion or contraction of output.

The market system exploits all possible trades between buyers and sellers. Up to the equilibrium quantity, buyers will pay more than suppliers require (those points on the demand curve lie above the supply curve). Beyond Q_1, buyers will not pay as much as suppliers need to produce more (those points on the supply curve lie above the demand curve). Again, in this regard the market can be called efficient.

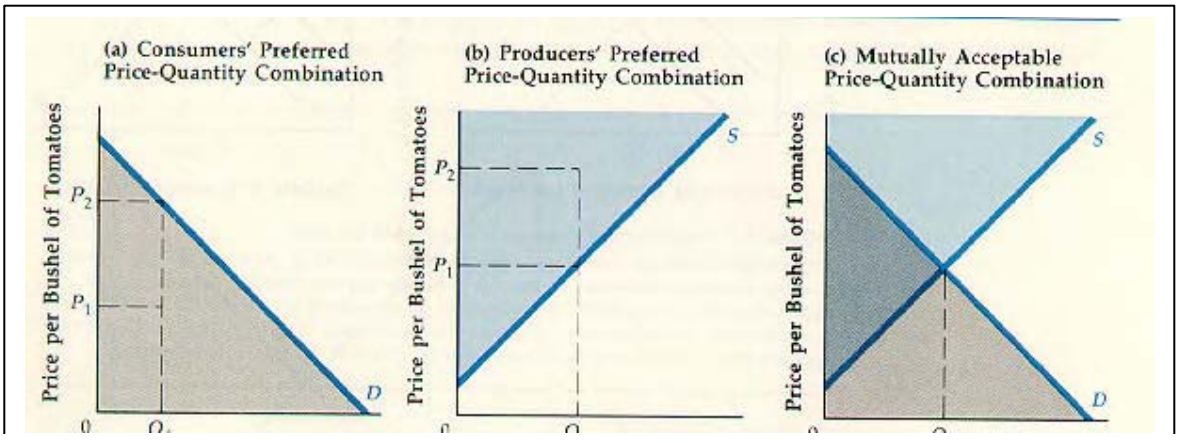


FIGURE 2.9 The Efficiency of the Competitive Market

Only those price-quantity combinations on or below the demand curve—panel (a)—are acceptable to buyers. Only those price-quantity combinations on or above the supply curve -- panel (b) -- are acceptable to producers. Those price-quantity combinations that are acceptable to both buyers and producers are shown in the darkest shaded area of panel (c). The competitive market is efficient in the sense that it results in output Q_1, the maximum output level acceptable to both buyers and producers.

Nonprice Competition

Markets in which suppliers compete solely in terms of price are relatively rare. Table salt is a relatively uniform commodity sold in a market in which price is an important competitive tool. Even producers of salt, however, compete in terms of real or imagined quality differences and the reputation and recognition of brand names. In most industries, competition is through a wide range of product features, such as quality or appearance, design, and durability. In general, competitors can be expected to choose the mix of features that gives them the greatest profit.

In fact, price competition is not always the best method of competition, not only because price reductions mean lower average revenues, but also because the reductions can be costly to communicate to consumers. Advertising is expensive, and consumers may not notice price reductions as readily as they do improvements in quality. Quality changes, furthermore, are not as readily duplicated as price changes. Consumers' preferences for quality over price should be reflected in the profitability of making such improvements. If consumers prefer a top-of-the-line calculator to a cheaper basic model, then producing the more sophisticated model could, depending on the cost of the extra features, be more profitable than producing the basic model and communicating its lower price to consumers.

If all consumers had exactly the same preferences—size, color, and so on—producers would presumably make uniform products and compete through price alone. For most products, however, people's preferences differ. To keep the analysis manageable, we will explore nonprice competition in terms of just one feature—product size. Suppose that in the market for television sets, consumer preferences are distributed along the continuum shown in Figure 2.10. The curve is bell shaped, indicating that most consumers are clustered in the middle of the distribution and want a middle-sized television. Fewer consumers want a giant screen or a mini-television.

Everything else being equal, the first producer to enter the market, Terrific TV, will probably offer a product that falls somewhere in the middle of the distribution—for example, at the in Figure 2.10. In this way, Terrific TV offers a product that reflects the preferences of the largest number of people. Furthermore, as long as there are no competitors, the firm can expect to pick up customers to the left and right of center. (Terrific TV's product may not come very

close to satisfying the wants of consumers who prefer a very large or very small television, but it is the only one available.) The more Terrific TV can meet the preferences of the greatest number of consumers, everything else being equal, the higher the price it can charge and the greater the profit it can make. (Because consumers value the product more highly, they will pay a higher price for it.)

The first few competitors that enter the market may also locate close to the center—in fact, several may virtually duplicate Terrific TV's product. These firms may conclude that they will enjoy a larger market by sharing the center with several competitors than by moving out into the wings of the distribution. They are probably right. Although they may be able to charge more for a giant screen or a mini-television that closely reflects some consumers' preferences, there are fewer potential customers for those products.

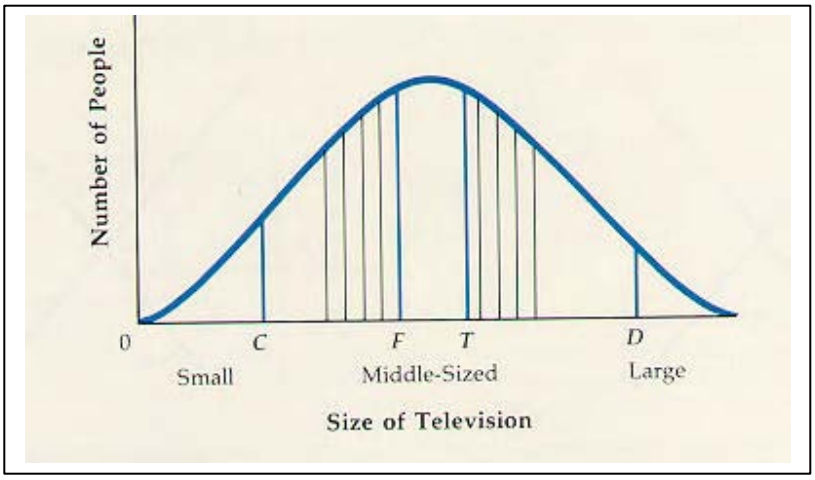


FIGURE 2.10 Consumer Preference in Television Size

Consumers differ in their wants, but most desire a medium-sized television. Only a few want very small or large televisions.

To illustrate, assume that competitor Fabulous Focus locates at F, close to T. It can then appeal to consumers on the left side of the curve because its product will reflect those consumers' preferences more closely than does Terrific TV's. Terrific TV can still appeal to consumers on the right half of the curve. If Fabulous Focus had located at *C*, however, it would have direct appeal only to consumers to the left of *C*, as well as to a few between *C* and *T*. Terrific TV would have appealed to more of the consumers on the left, between *C* and *T*, than in the first case. In short, Fabulous Focus has a larger potential market at *F* than at *C*.

However, as more competitors move into the market, the center will become so crowded that new competitors will find it advantageous to move away from the center, to *C* or *D*. At those points the market will not be as large as it is in the center, but competition will be

less intense. If producers do not have to compete directly with as many competitors, they can charge higher prices. How far out into the wings they move will depend on the tradeoffs they must make between the number of customers they can appeal to and the price they can charge.

Like price reductions, the movement of competitors into the wings of the distribution benefits consumers whose tastes differ from those of the people in the middle. These atypical consumers now have a product that comes closer to or even directly reflects their preferences.

Our discussion has assumed free entry into the market. If entry is restricted by monopoly of a strategic resource or by government regulation, the variety of products offered will not be as great as in an open, competitive market. If there are only two or three competitors in a market, everything else being equal, we would expect them to cluster in the middle of a bell-shaped distribution. That tendency has been seen in the past in the broadcasting industry, when the number of television stations permitted in a given geographical area was strictly regulated by the Federal Communications Commission. Not surprisingly, stations carried programs that appealed predominantly to a mass audience—that is, to the middle of the distribution of television watchers. The Public Broadcasting System, PBS, was organized by the government partly to provide programs with less than mass appeal to satisfy viewers on the outer sections of the curve. When cable television emerged and programs became more varied, the prior justification for PBS subsidies became more debatable.

Even with free market entry, product variety depends on the cost of production and the prices people will pay for variations. Magazine and newsstand operators would behave very much like past television managers if they could carry only two or three magazines. They would choose *Newsweek* or some other magazine that appeals to the largest number of people. Most motel operators, for instance, have room for only a very small newsstand, and so they tend to carry the mass-circulation weeklies and monthlies.

For their own reasons, consumers may also prefer such a compromise. Although they may desire a product that perfectly reflects their tastes, they may buy a product that is not perfectly suitable if they can get it at a lower price. Producers can offer such a product at a lower price because of the economies of scale gained from selling to a large market. For example, most students take pre-designed classes in large lecture halls instead of private tutorials. They do so largely because the mass lecture, although perhaps less effective, is substantially cheaper than tutorials. In a market that is open to entry, producers will take advantage of such opportunities.

If producers in one part of a distribution attempt to charge a higher price than necessary, other producers can move into that segment of the market and push the price down; or consumers can switch to other products. In this way, an optimal variety of products will eventually emerge in a free, reasonably competitive market. Thus the argument for a free market is an argument for the optimal product mix. Without freedom of entry, we cannot tell whether it is possible to improve on the existing combination of products. A free, competitive market gives rival firms a chance to better that combination. The case for the free market becomes even stronger when we recognize that market conditions—and therefore the optimal product mix—are constantly changing.

Competition in the Short run and the Long Run

One of the best examples of the workings of both price and nonprice competition is the market for hand calculators. Since the first model was introduced in the United States in 1969, the growth in sales, advancement in technology and design, the decline in prices in this market have been spectacular. The early calculators were simple—some did not even have a division key—and bulky by today's standards. By 1976 they had shrunk from the size of a large paperback book to a tiny two by three-and-a-half inches for one model, and sales exceeded 16 million.

While quality improved, prices fell. The first calculator, which Hewlett-Packard sold for $395, had an eight-digit display and performed only four basic functions—addition, subtraction, division, and multiplication. By December 1971 Bowmar was offering an eight-digit, four-function model for $240. The next year, in an attempt to maintain its high prices, Hewlett-Packard introduced a sophisticated model that could perform many more functions, still for $395. By the end of the year, Bowmar, Sears, and other firms had broken the $100 barrier, and firms were offering built-in memories, AC adapters, and 1,500-hour batteries to shore up prices. At the year's end, Casio announced a basic model for $59.95.

In 1973 prices continued to fall. By the end of the year, National Semiconductor was offering a six-digit, four-function model for $29.95, and Hewlett-Packard had lowered the price of its special model by $100 and added extra features. In 1974, six-digit, four-function models sold for as little as $16.95. Eight-digit models that would have sold for over $300 three or four years earlier carried price tags of $19.95. By 1976 consumers could buy a six-digit model for just $6.95. All this happened during a period when prices in general rose at a rate unprecedented in the United States during peacetime. Thus the relative prices of calculators fell by even more than their dramatic price reductions suggest.

Yet the drop in the price of calculators was to be expected. Although the high prices of the first calculators partly reflected high production costs, they also brought high profits and tempted many other firms into the industry. These new firms duplicated and then improved the existing technology and increased their productivity in order to beat the competition or avoid being beaten themselves. Firms unwilling to move with the competition quickly lost their share of the market.

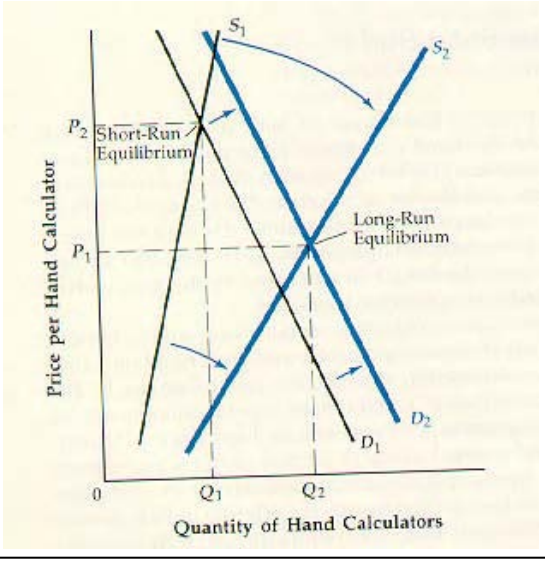


FIGURE 2.11 Long-Run Market for Calculators

With supply and demand for calculators at D_1 and S_1, the short-run equilibrium price and quantity will be P_2 and Q_1. As existing firms expand production and new firms enter the industry, the supply curve shifts to S_2. Simultaneously, an increase in consumer awareness of the product shifts the demand curve to D_2. The resulting long-run equilibrium price and quantity are P_1 and Q_2.

The increase in competition in the calculator market can be represented visually with supply and demand curves. Such an analysis permits us to observe long-run changes in market equilibrium. Given the limited technology and the small number of firms producing calculators in 1969, as well as restricted demand for this new product, let us assume that the supply and demand curves were initially S_1 and D_1 in Figure 2.12. The initial equilibrium price would then be P_2 and Q_1. This is the short-run equilibrium. **Short-run equilibrium** is the price-quantity combination that will exist as long as producers do not have time to change their production facilities (or some resource that is fixed in the short run).

Short-run equilibrium did not last long. In the years following 1969, firms expanded production, building new plants and converting facilities that had been producing other small electronic devices. Economies of scale resulted, and technological breakthrough lowered the cost of production still further. Several $150 circuits were reduced to very small $2 chips. The increased supply shifted the supply curve to the right, from S_1 to S_2 (see Figure 2.12). Meanwhile, because of advertising and word of mouth, people became familiar with the product and market demand increased, shifting the demand curve from $D_1$to D_2. Because supply increased more than demand, the price fell from P_2 to P_1, and quantity rose from Q_1 to Q_2. The new equilibrium price and quantity, P_1 and Q_2, marked the new long-run market equilibrium. **Long-run equilibrium** is the price-quantity combination that will exist after firms have had time to change their production facilities (or some other resource that is fixed in the short run).

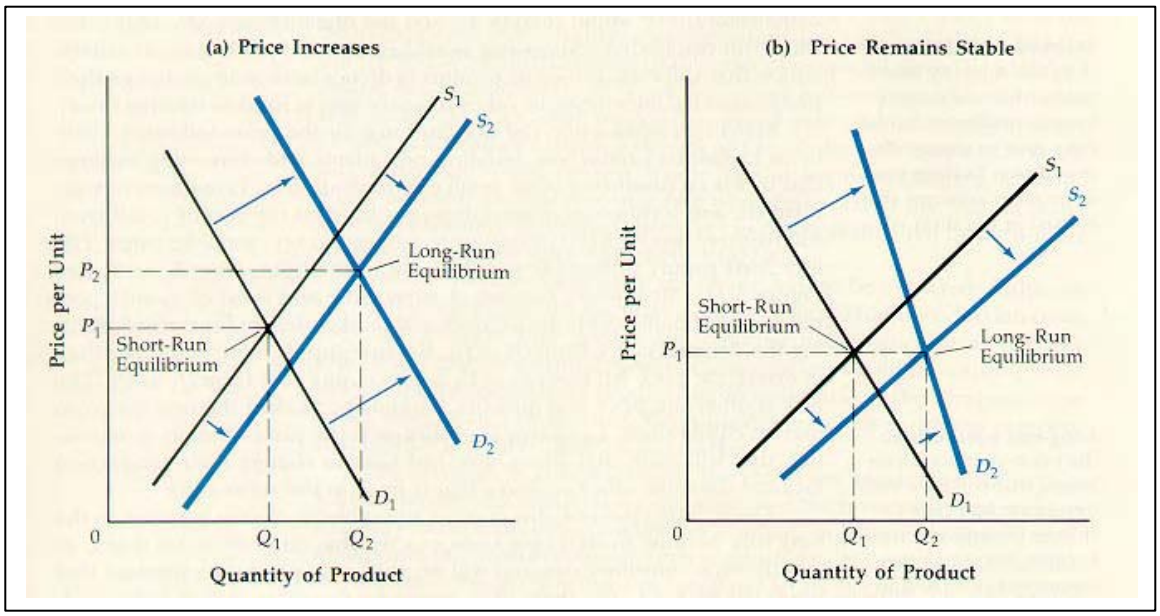


FIGURE 2.12 Prices in the Long Run

If demand increases more than supply, the price will rise along with the quantity sold—panel (a). If supply keeps up with demand, however, the price will remain the same even though the quantity sold increases—panel (b).

The market does not always move smoothly from the short run to the long run. Because firms do not know exactly what other firms are doing, or exactly what consumer demand will be, they may produce a product that cannot be sold at a price that will cover product costs. In fact, in the mid-1970s prices fell enough that several companies were losing money. Long-run improvements sometimes come at the expense of short-run losses.

In this example, a long-run market adjustment causes a drop in price (because supply increased more than demand). The opposite can occur: demand can increase more than supply, causing a rise in the price and the quantity produced. In Figure 2.12(a), when the supply curve shifts to S_2 and the demand curve shifts to D_2, price increases from P_1 to P_2 and quantity produced rises from Q_1 to Q_2. Supply and demand may also adjust so that price remains constant while quantity increases (Figure 2.12(b)).

Shortcomings of Competitive Markets

Although the competitive markets may promote long-run improvements in product prices, quality, and output levels, it has deficiencies, and we must note several before closing. (Market deficiencies will be discussed further in later chapters.)

First, the competitive market process can be quite efficient because production is maximized. Consumer demand, however, depends on the way income is distributed. If market forces or government programs distort income distribution, the demand for goods and services will also be distorted. If, for example, income is concentrated in the hands of a few, the demand for luxury items will be high, but the demand for household appliances and new housing will be low. In such a situation, the results of competition may be efficient in a strict economic sense, but whether these results are socially desirable is a matter of values—of normative, rather than positive, economics.

Second, the outcome of competition will not be efficient to the extent that production costs are imposed on people who do not consume a product. People whose house paint peels because of industrial pollution bear a portion of the offending firm's production cost, whether or not they buy its product. At the same time, the price consumers pay for the product is lower than it would be if the producer incurred all costs, including pollution costs. Because of the low price, consumers will buy more than the efficient quantity. In a sense, this is an example of overproduction. Because all the costs of production have not been included in the producer's cost calculations, the price is artificially low.

Third, in a free market, competition can promote socially undesirable products or services. A competitive market in an addictive drug like alcohol or heroin can lead to lower prices and greater quantities consumed -- and thus an increase in social problems associated

with addition. Competition can be desirable only when it promotes the production of things people consider beneficial, but what is beneficial is a matter of values.

Fourth, opponents of the market system contend that competition sometimes leads to "product proliferation" -- too many versions of essentially the same product, such as aspirin—and to waste in production and advertisement. Because so many types of the same product are available, production of each takes place on a very small scale, and no plant is fully utilized. This may be true. The validity of this objection, however, hinges on whether the range of choice in products compensates for the inefficiencies in production. The question is whether firms should be forced to standardize their products and to compete solely in terms of price. What about people who want something different from the standard product?

Fifth, unscrupulous competitors can take advantage of customers' ignorance. A competitor may employ unethical techniques, such a circulating false information about rivals or using bait-and-switch promotional tactics (advertising very low-priced, low-quality products to attract customers and to switch them to higher-priced products when they get into the store). Competition can control some of these abuses. For instance, competitors will generally let consumers know when their rivals are misrepresenting their products. Still, fraudulent sellers can move from one market to another, keeping one step ahead of their reputations.

MANAGER'S CORNER: **Paying Above-Market Wages**[5]

This chapter has been about how "markets" do things like set product prices and production levels through the forces of competition. However, markets don't operate by themselves. Real live people are involved who sometimes seem to do things that defy conventional market explanation. Take, for example, Henry Ford who is remembered for his organizational inventiveness (the assembly line) and for his presumption that he could ignore the wishes of his customers (as in his claim that he was willing to give buyers any color car they wanted so long as it was black!). However, he outdid himself when it came to workers; he *seemed* to want to deny the control of the market when it came to setting his workers' wages. Did he really?

In 1914, he stunned his board of directors by proposing to raise his workers' wages to $3 a day, a third higher than the going wage ($2.20 a day) in the Detroit automobile industry at the time. When one of his board members wondered out loud why he was not considering giving workers even more, a wage of $4 or $5 a day, Ford quickly agreed to go to $5, more than twice the prevailing market wage. Why?

---

[5] Reprinted from Richard B. McKenzie and Dwight R. Lee, Managing Through Incentives (New York: Oxford University Press, 1998), chap. 6.

An answer to why Ford paid more than the prevailing wage won't be found on the pages of standard economics textbooks.[6] In those texts, wages are determined by market conditions, namely, the forces of supply and demand, and demand and supply (often depicted by intersecting lines on a graph) are locked in place, that is, are not affected by how, or how much, workers are paid. The supply of labor is determined by what workers are willing to do, while the demand for labor is determined by the combined forces of worker productivity and the prices that can be charged for what the workers produce. The curves are more or less stationary (at least in the way they are presented), certainly not subject to manipulation by employers and their policies.

In the competitive framework, the "market wage" will settle where the market clears, or where the number of workers who are demanded by employers exactly equals the number of workers who are willing to work. And, once more, no profit-hungry employer (at least in the textbook discussions) would ever pay above (or below) market. For that matter, in standard textbooks, employers in competitive markets are *unable* to pay anything other than the market wage, given competition. If employers ever tried to pay more, they could be underpriced by other producers who paid less, the market wage. If employers paid below market, they would not be able to hire employees and would be left without products to sell.

There are two problems with that perspective from the point of view of this book. First, we don't wish to assume away the problem of policy choices. On the contrary, we want to discuss how policies might affect worker productivity, or how employers might achieve maximum productivity from workers. We seek a rationale for Ford's dramatic wage move, if there is one to be found. In doing so, we don't deny that productivity affects worker wages, which is a well-established theoretical proposition in economics. What we insist on is that the reverse is also true -- worker wages affect productivity -- for very good economic reasons.

Second, a problem with standard market theory is that there is a lot of real-world experience that does not seem to fit the simple supply and demand model. Granted, the standard model is highly useful for discussing how wages might change with movements in the forces of supply and demand. From that framework, we can appreciate, for example, why wages move up when the labor demand increases (which can be attributable to productivity and/or price increases). At the same time, many employers have followed Ford's lead and have paid more than market wages. All one has to do to check out that claim is to watch how many workers put in applications when a plant announces it is hiring. Sometimes, the lines stretch for blocks from the plant door. When the departments of history or English in our universities have an open professorship, the departments can expect a hundred or more qualified applicants. The U.S. Postal Service regularly receives far more applications for its carrier jobs than it has jobs available. When Boeing came to Los Angeles in late 1996 to hire workers, the line-up at the work fair stretched for blocks down the street; the end, in fact, could not be seen from the door. These queues cannot be explained by market clearing wages.

---

[6]Our discussion on the Ford pay increase is heavily dependent on a book by Stephen Meyer, The Five-Dollar Day: Labor, Management, and Social Control in the Ford Motor Company, 1908-1921 (Albany, N.Y.: State University of New York Press, 1981).

Consider the persistence of unemployment. The traditional view of labor markets would predict that the wage should be expected to fall until the market clears and the only evident unemployment should be transitory, encompassing people who are not working because they are between jobs or are looking for jobs. But "involuntary unemployment" abounds and persists, which must be attributable, albeit partially, to paying workers "too much" (or above the market-clearing wage rate).

We don't pretend to provide a complete explanation for "overpaying" workers here. It may be that employers overpay their workers for some psychological reasons. Overpaying workers might make the employers feel good about themselves and their employees, which can show up in greater loyalty, longer job tenure, and harder and more dedicated work. The above-market wages may also remove workers' financial strains, leaving them with fewer problems at home and more energy to devote themselves to their jobs. While we think these can be important considerations, we prefer to look for other reasons, mainly as a means of improving incentives for workers to do as the employer wants.

As it turns out, Henry Ford was not offering his workers something extra for nothing in return. He wanted to "overpay" his workers primarily because he could then demand more of them. He could work them harder and longer, and he did. He could also be more selective in the people he hired, which could be a boon to all Ford workers. Workers could reason that they would be working with more highly qualified cohorts, all of whom would be forced to devote themselves to their jobs more energetically and productively. Some, if not all, of the wage would be returned in the form of greater production and sales and even greater job security for workers. But there were other benefits for Ford.

When workers are paid exactly their market wage, there is little cost to quitting. A worker making his market (or opportunity) wage can simply drop his job and move on to the next job with no loss in income. And, as was the case, Ford's workers were quitting with great frequency. In 1913, Ford had an employee turnover rate of 370 percent! That year, the company had to hire 52,000 workers to maintain a workforce of 13,600 workers.

The company estimated that hiring a worker costs from $35 to $70, and even then they were hard to control. For example, before the pay raise, the absentee rate at Ford was 10 percent. Workers could stay home from work, more or less when they wanted, with virtually no threat of penalty. Given that they were being paid market wage, the cost of their absenteeism was low to the workers. In effect, workers were buying a lot of absent days from work. It was a bargain. They could reason that if they were only receiving the "market wage rate," then that wage rate could be replaced elsewhere if they were ever fired for misbehaving on the job.

At any one time, most workers were new at their job. Shirking was rampant. Ford complained that "the undirected worker spends more time walking about for tools and material than he does working; he gets small pay because pedestrianism is not a highly paid line." In order to control workers, the company figured that the firm had to create some buffer between itself and the fluidity of a "perfectly" functioning labor market.

The nearly $3 Ford paid above the market was, in effect, a premium he had to pay in order to enforce the strict rules for employment eligibility he imposed. Ford's so-called Sociology Department was staffed by investigators who, after the pay hike, made frequent home visits and checked into workers' savings plan, marital happiness, alcohol use, and moral conduct, as well as their work habits on the job. He was effectively paying for the right to make those checks, and he made the checks in part because he thought they were the right thing to do, but also because the checks would lead to more productive workers.

Ford was also paying for obedience. He is quoted as saying after the pay hike, "I have a thousand men who if I say 'Be at the northeast corner of the building at 4 a.m.' will be there at 4 a.m. That's what we want -- obedience."[7] Whether he got obedience or allegiance may be disputed. What is not disputable is that he got dramatic results. In 1915, the turnover rate was 16 percent -- down from 370 percent -- and productivity increased about 50 percent.

It should be pointed out that control over workers is only part of the problem. Even if a boss has total control, there must be some way of knowing what employees should be doing to maximize their contribution to the firm. That wasn't a difficult problem for Ford. On the assembly line, it was obvious what Ford wanted his workers to do, and it was relatively easy to spot shirkers. According to David Halberstam in his book The Reckoning, there was small chance for the shirker to prosper in the Ford plant. After the plant was mechanized and the $5-a-day policy was implemented, foremen were chosen largely for physical strength. According to Halberstam, "If a worker seemed to be loitering, the foreman simply knocked him down."[8] Given that the high wage attracted many applicants, Ford's workers simply had to put up with the abuse and threat of abuse, or be replaced. The line outside the employment office was a strong signal to workers.

Of course, this type of heavy-handed control doesn't work in every work environment. When productivity requires that workers possess a lot of specialized knowledge that they must exercise creatively or in response to changing situations, heavy-handed enforcement tactics may not work effectively. Indeed, the threat can undermine creativity and productivity. How is a manager to know whether a research chemist, a creator of software, or a manager, is behaving in ways that make the best use of his talents in promoting the objectives of the firm? Do you knock them down if they gaze out the window? Managers typically provide a subtler incentive program than a high daily salary and a tough foreman. The big problem is controlling employees who have expertise you lack.

---

[7]David Halberstam, The Reckoning (New York: Avon Books, 1986), p. 94.
[8]Ibid.

One way to inspire effort from those who can't be monitored directly on a daily basis is to "overpay" workers, and ensure that they suffer a cost in the event that their performance, as measured over time, is not adequate. The "overpayment" gives workers a reason to avoid being fired or demoted for such reasons as lack of performance and excessive shirking. Even when shirking is hard to detect, the threat of losing a well-paying job can be sufficient to motivate diligent effort.[9]

Many workers are in positions of responsibility, meaning that they have control over firm resources (real and financial) that they typically use with discretion -- and can also misuse, or appropriate for their own uses. Their actions are also difficult to monitor. Misuse of funds may only infrequently be discovered. How should such employees be paid? More than likely, they should be "overpaid." That is, they should be paid more than their market wage as a way of imposing a cost on them if their misuse of funds -- especially, their dishonesty -- is ever uncovered. The expected lost "excess wages" must exceed the potential (discounted) value of the misused funds. The less likely the employees are to be found out, the greater the overpayment must be in order for the cost to be controlling.

For example, assume a person receives a wage premium of $100 because he or she is in a position of trust and has control over firm resources. If the person can expect to be discovered one out of every ten times he steals firm property, at which point he will lose his job and his wage premium, the employee would assess the expected cost of theft at $10 per instance. The person who expected to be caught much less frequently, say, one out of every 100 times, would assess the expected cost at $1. To balance the expected cost in the two instances, the wage premium would have to be higher in the latter case (or $1,000). Of course, it naturally follows that, given the probability of being caught, the more a person can steal from the firm (or the more firm resources the employee can misuse or misdirect), the greater must the wage premium be to have the same deterrent effect.

Why do managers of branch banks make more than bank tellers? One reason is that the managers' talents are scarcer than tellers' are. That is a point frequently drawn from standard labor-market theorizing. We add here two additional factors: First, the manager is very likely in a position to misuse, or just steal, more firm resources than is each individual teller. Second, the manager's actions are less likely to be discovered than the teller's. The manager usually has more discretion than each teller does, and the manager has one less level of supervision.

---

[9] See J. Bulow and Lawrence Summers, "A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination and Keynesian Unemployment," Journal of Labor Economics, vol. 4 (no. 3, July 1986), pp. 376-414; and C. Shapiro and Joseph Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," American Economic Review, vol. 74, no. 3 (June 1984), pp. 433-444. So-called "equity theory," based in psychology, suggests that worker over-payment can lead to greater performance because the overpaid workers perceive an inequity in pay among their relevant peers. As a consequence, they seek to redress the overpayment by working longer and harder. Of course, the theory also suggests that underpaid workers will respond by working less diligently and putting in less time. See Edward E. Lawler, III, "Equity Theory as a Predictor of Productivity and Work Quality," Psychological Bulletin, vol. 70 (no. 6, December 1968), pp. 596-610.

Why does pay escalate with rank within organizations? There are myriad reasons, several of which will be covered later. We suggest here that as managers move up the corporate ladder, they typically acquire more and more responsibility, gain more discretion over more firm resources, and have more opportunities to misuse firm resources. In order to deter the misuse of firm resources, the firm needs to increase the threat of penalty for any misuse, which implies a higher and higher wage premium for each step on the corporate ladder.

Workers in the bowels of their corporations often feel that the people in the executive suite are drastically "overpaid," given that their pay appears to be out of line with what they do. To a degree, the workers are right. People in the executive suite are often paid a premium simply to deter them from misusing the powers of the executive suite. The workers should not necessarily resent the overpayments. The overpayments may be the most efficient way available for making sure that firm resources are used efficiently. To the extent that the overpayments work, the jobs of people at the bottom of the corporate ladder can be more productive, better paying, and more secure.

We have not covered all possible reasons workers are not paid strictly as suggested by simple supply and demand curve analysis. Nevertheless, the Ford case permits us to make two general points: First, moving decisions away from the impersonal forces of the marketplace and into the more personal forces inside a firm, with long-term relational contracts, can increase efficiency by reducing transaction costs. And, second, the decisions made on how the firms organize their "overpayments" can have important consequences for the efficiency of production because workers can have a greater incentive to invest "sweat equity" in their firms and to become more productive. The firm that gets the "overpayment" right (and exactly what it should be cannot be settled in theory) can gain a competitive advantage over rivals. Apparently, Ford secured an important advantage by going, in a sense, "off market."

Should workers accept "overpayment"? Better yet, is a greater overpayment always better for workers? The natural tendency is to answer with a firm, "Yes!" Well, we think a more cautious answer is in order, "Maybe" or, again, "It depends." Workers would be well advised to carefully assess what is expected of them, immediately and down the road. High pay means employers can make greater demands -- in terms of the scope and intensity of work assignments -- on their employees. This is because of the cost they will bear if they do not consent to the demands.

Clearly, workers should expect that their employers will demand value equal to, if not above, the wage payments, and workers should consider whether they contribute as much to their firms' coffers as they take. Otherwise, their job tenure may be tenuous. The value of a job is ultimately equal to how much the workers can expect to earn over time, appropriately adjusted for the fact that future payments are not worth as much to workers as current ones are and for the fact that uncertain payments are not worth as much as certain payments. A high paying job that is lost almost immediately for inadequate performance may be a poor deal for employees.

To make this point with focus in our classes, we have often told our MBA students that they are unlikely to be offered upon graduation salaries at the high end of the executive level.

However, if by some chance they were offered such a salary -- say, $250,000 a year -- they should seriously consider turning it down. We suggest that most should probably consider jobs with annual salaries more in the range of $50,000 to $70,000, something close to whatever is the going market wage for their graduate school cohorts. Our students are generally startled by our brazen suggestion.

Why should any sane person turn down such a lucrative offer, if a sane employer tendered it? An answer is not all that mysterious. Unless a new graduate is able and willing to return $250,000 a year in value, he or she would be unlikely to retain such a high paying job for very long. The person who quickly fails at a high salary can end up doing far worse than the person who begins her career by succeeding at a more modest salary.

The point that emerges from such a discussion and needs to be remembered is that the actual extent of the "overpayment" will not be determined solely by employers, as was true with Ford in 1913. Employees will also have a say. They have an interest in limiting the overpayment in order to limit the demands placed on them and to increase their job security. That is to say, the extent of the "overpayment" is, itself, determined by negotiation, if not market forces, with the wage pressures not always in the way expected. The pay negotiations can involve the workers pressing for a lower overpayment while the employer presses for the opposite.

Along this line, we have seriously suggested in another book (but with little hope of being taken seriously by political operatives) that members of Congress should not have control over their own pay.[10] By restricting their overpayment, they thwart the competition for their jobs and increase their job security -- and the current value of being in Congress. As opposed to cutting their pay in order to reduce the net value of being in Congress, we suggest it might be a wiser course to increase the members' pay rather dramatically to, say, half a million a year. That could increase the competition in congressional races, increase the quality of candidates who run, and undercut the job security for members of Congress. At the same time, the higher pay could make members far more responsive to voter interests than the current pay does by imposing formula driven reductions in their pay if deficits or inflation exceed specified levels.

Firms might also "overpay" their workers because they have "underpaid" their workers early in their careers. The "overpayments" are not so much "excess payments" as they are "repayments" of wages forgone early in the workers' careers. Of course, the workers would not likely forgo wages unless they expected their delayed overpayments to include interest on the wages forgone. So, the delayed overpayments must exceed underpayments by the applicable interest market interest rate. In such cases, the firms are effectively using their workers as sources of capital. The workers themselves become venture capitalist of an important kind.

Why would firms do that? Some new firms must do it just to get started. They don't have access to all of the capital they need in their early years, given their product or service has

[10]Dwight R. Lee and Richard B. McKenzie, Regulating Government: A Preface to Constitutional Economics (Lexington, Mass.: Lexington Books, 1987), pp. 157-162.

not been proven. They must ask their workers to invest "sweat equity," which is equal to the difference between what the workers could make in their respective labor markets and what they are paid by their firms. The underpayments not only extend the sources of capital to the firm, but they also give the workers a strong stake in the future of the firm, which can make the workers work all the harder to make the firm's future a prosperous one. The up-front underpayments can make the firm more profitable and increase its odds of survival, which can be a benefit to the workers as well as owners. Of course, this is one reason many young workers are willing to accept employment in firms that are just starting out. Young workers often have a limited financial base from which to make investments; they do, however, have their time and energy to invest.

Underpayments to workers coupled with later overpayments can also be seen as a means by which managers can enhance the incentives workers have to become more productive. If workers are underpaid when they start, their rewards can be hiked later by more than otherwise to account for productivity improvements. These hikes can continue – and must continue -- until the workers are effectively overpaid later in their careers (or else the workers would not have accepted the underpayments earlier in their careers). However, managers must understand that they must be able to *commit* themselves to the overpayments and that there must be some end to the overpayments.

Not too many years ago, firms regularly required their workers to retire at age 65. Retirement was ritualistic for managers. Shortly after a manager had his or her sixty-fifth birthday, someone would organize a dinner at which the manager would be given a gold watch and a plaque for venerable service and then be shown to the door with one last pleasant goodbye.

Why would a firm impose a mandatory retirement age on its workers? Such a policy seems truly bizarre, given that most companies are intent on making as much money as they can. Often the workers forced to retire are some of the more productive in the firm, simply because they have more experience with the firm and its customer and supplier networks.

While we acknowledge that mandatory retirement may appear mistaken, particularly in the case of highly productive employees, we think that for many companies a mandatory retirement policy makes good business sense – when they have been "overpaying" their workers for sometime. (Otherwise, we would be hard pressed to explain why such policies would survive and would need to be outlawed.) To lay out that logic, we must take a detour into an analysis of the way workers, who come under mandatory retirement policies, are paid throughout their careers.

Paying market wages, or exactly what workers are worth at every stage in the worker's career, does not always maximize worker incomes. That was a central point of the discussion to this point. We extend that discussion here by showing how the manipulation of a worker's *career* wage structure, or earnings path over time, can actually raise worker productivity and lifetime income. However, as will also be shown, when worker wages diverge from their value over the course of their careers, mandatory retirement is a necessary component of the labor contract.[11]

Suppose that a worker goes to work for Apex, Inc. and is paid exactly what she is worth at every point in time. Assume she can expect to have a modest productivity improvement over the course of a thirty-year career, described by the slightly upward sloping line A in Figure 2.13. If her income follows her productivity, her salaries will rise in line with the slope of line A. In year Y_1, the worker's annual income will be I1; in year Y_2, it will be I_2, and so forth.

Is there a way by which management can restructure the worker's income path and, at the same, enable both the workers and the firm to gain? No matter what else is done, management must clearly pay the worker an amount equal at least to what he or she is worth *over the course of her career*. Otherwise, the worker would not stay with the company. The worker would exit the firm, moving to secure the available higher career income. However, management need not pay each year an amount equal to the income points represented on line A. Management could pay the worker less than she is worth for awhile so long as management is willing to compensate by overpaying her later.

For example, suppose that management charts a career pay path given by line B, which implies that up until year Y_3, the workers are paid less than they are worth, with the extent of the underpayment equaling the shaded area between the origin and *Y*3. However, the workers would be compensated for what amounts to an investment in the firm by an overpayment after year Y_3, with the extent of the overpayment equal to the shaded area above line *A* after Y_3.

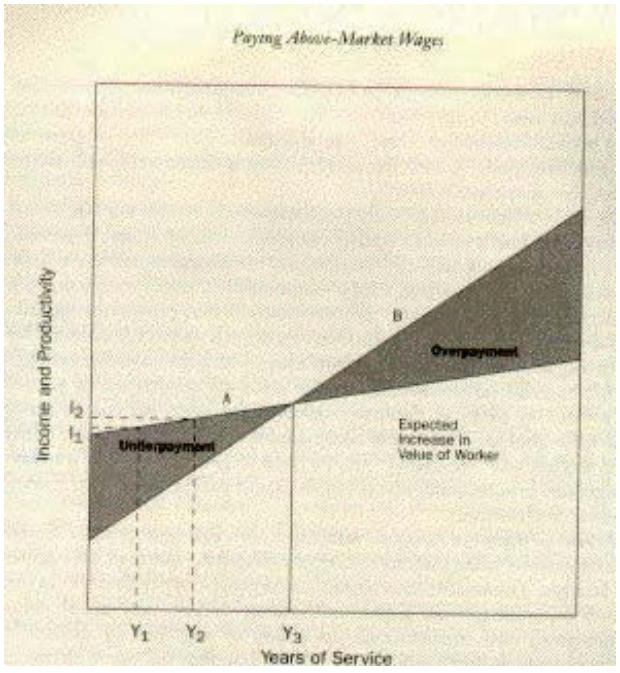


FIGURE 2.13 Twisted Pay Scale

The worker expects his productivity to rise alone line *A* with years of service. If she starts work with less pay that she could earn elsewhere, then her career pay path could follow line *B*, representing greater increases in pay with time and greater productivity.

[11]For the analysis presented here, we are indebted Lazear [Edward Lazear, "Why Is There Mandatory (December 1979), pp. 1261-1284].

Are the firm and worker likely to be better off? Notice that the actual proposed pay line *B* is much steeper than line *A*, which, again, represents the worker's income path in the absence of management's intentional twisting of the pay structure. The greatest angle of line *B* means that the worker's income rises by more than warranted by the year-to-year increases in her productivity. This implies that the worker has a greater incentive to actually do what management wants done, which is increase productivity. This is the case because the worker gets a disproportionately greater reward for any given productivity improvement. The increase in productivity can translate into greater firm revenue, which can be shared between the workers, management, and owners.

Would workers ever want to work for a firm that intentionally underpays its workers when they are young or just starting out with the company? You bet. The workers can reason that everyone in the firm will have a greater incentive to work harder and smarter. Hence, they can all enjoy higher prospective incomes over the course of their careers. Normally, commentaries on worker pay implicitly assume that the pay structure is what management imposes on workers. Seen from the perspective of the economic realities of what is available for distribution to all workers in a firm, we could just as easily reason that the kind of pay structure represented by line B is what the workers would encourage management to adopt. Actually, the twisting of the pay structure is nothing more than an innovative way for managers to increase the money they make off their workers while also increasing the money workers are able to make off their firms. In short, it is a mutually beneficial deal, something of a "free good," in the sense that more is available for everyone.

If twisting the pay structure is such a good idea, why isn't it observed more often than it is in industry? Perhaps some variant of twisted pay schedules is more widely used than thought, primarily because they are not identified as such. Public and private universities are notorious for making their assistant professors work harder than full professors who have tenure and far more pay. Large private firms, like General Motors and IBM, appear to have pay structures that are more like line B than line A. However, millions of firms appear to be unwilling or unable to move away from a pay structure like line A.

One of the problems with line B is that young workers must accept a cut in pay for a promise of greater pay in the future -- and the pay later on must exceed what the workers can get elsewhere *and*, what is crucial to workers, more than what their firm would have to pay if they simply hired replacement workers at the going market wage. Obviously, the workers take

a considerable risk that their firm will not live up to its promise by deciding not to raise their pay later to points above their market wage or, what is worse, fire them.

Needless to say, the firm must be able to make a *credible commitment* to its workers that it will live up to its part of the bargain, the *quo* in the *quid pro quo*. Truly *credible commitments* require that the firm be able to demonstrate a capacity and inclination to do what it says it will do. The firm must be believable by those who make the early wage concessions. Many firms are not going to be able to twist their pay structures, and gain the productivity improvements, because they are new, maybe small, with a shaky financial base and an uncertain future. New firms have little history for workers to assess the value of their firms' commitments. Small firms are often short-lived firms. Financially shaky firms, especially those which suffer from problems of insolvency or illiquidity, will unlikely be able to garner the trust of their workers. Firms that are in highly fluid, ever-changing and competitive markets, will also be unlikely candidates for being able to twist their pay structures. They all will tend to have to pay workers their market worth, or even a premium to accommodate the risks the workers must accept when the company's existence is in doubt.

What firms are most likely to twist their pay structures? Ones that have been established for some time, have a degree of financial and market stability, have some monopoly power -- and have proven by their actions that their word is their bond. To prove the latter, firms cannot simply go willy-nilly about dismissing workers or cutting their pay when they find cheaper replacements. To do so would be an undermining of their credibility with their workers.

We can't be too precise in identifying the types of firms that can twist their pay structures for the simple reason that there can be extenuating circumstances. For example, we can imagine some unproved up-start companies would be able to pay their workers below market wages. Indeed, they may have to do so simply because they do not have the requisite cash flow early in their development. New firms often ask, or demand, that their workers provide "sweat equity" in their firms through the acceptance of below-market wages, but always with the expectation that their investment will pay off. Which new firms are likely to be able to do this? We suspect that firms with new products that represent a substantial improvement over established products would be good candidates. The likely success of the new product gives a form of base-line credibility to firm owner commitments that they intend -- and can -- repay the "sweat equity" later. Indeed, the greater the improvement the new product represents, the more likely the firm can make the repayment, and do so in an expeditious manner, and the more likely the workers will accept below-market wages to start. The very fact that the product is a substantial improvement increases the likelihood of the firm's eventual success for two reasons. The first reason is widely recognized: a product that represents a substantial improvement will likely attract considerable consumer attention. The second reason is less obvious: the firm can delay its wage payments, using its scarce cash flow in its initial stages of production for other things, such as quality control, distribution, and promotion. The firm gets capital -- sweat equity -- from an unheralded source, workers. The workers' investment of their sweat equity can enhance the firms' survival chances and, thereby, even lower the interest rate that the firms must pay on their debt (because the debt is more secure).

Of course, there are times when firms must break with their past commitments. For example, if a firm, which was once insulated from foreign competition, must all of a sudden confront more cost-effective foreign competitors in domestic markets (because, say, transportation costs have been lowered), then the firm may have to break with its commitments to overpay workers late in their careers. If they don't, the competition will simply pay people the going market wage and erode the markets of those firms who continue to overpay their older workers. Without question, many older American workers, for example, middle managers in the automobile industry, have hard feelings about the advent of the "global marketplace." They may have suffered through years of hard work at below-market wages in the belief that they would be able, later in their careers, to slack off and still see their wages rise further and further above market. The advent of global competition, however, has undercut the capacity of many American firms to fulfill their part of an implied bargain with their workers.

Even though they may have hard feelings, it does not follow that the workers would want their firms to try to hold to their prior agreements. Many workers understand that their wages can be higher *than they otherwise would be* if their firms kept their prior agreement. Without the reneging, the firm might fold. In a sense, the workers made an investment in the firm through their lower wages, and the investment didn't pay off as much as expected. However, we hasten to add that some American workers have probably been burned by firms that have used changing market conditions as an *excuse* to

break with their commitments or that have sold their firms to buyers who felt no compulsion to hold to the original owners' prior commitments.[12]

The answer to the question central to this section, "Why does mandatory retirement exist?" can now be provided, at least partially. Mandatory retirement at, say, 65 or 70 may be instituted for any number of plausible reasons. It might be introduced simply to move out workers who have become mentally or physically impaired. Perhaps, in some ideal world, the policy should not, for this reason, be applied to everyone. After all, many older workers are in the midst of their more productive years, because of their accumulated experience and wisdom, when they are in their sixties and seventies. However, it may still be a reasonable *rule* because its application to *all* workers may mean that *on average*, by applying the policy without exception, the firm is more efficient and profitable.

However, the *expected* fitness of workers at the time of retirement is simply not the only likely issue at stake. We see mandatory retirement as we see all employment rules, as a part of what is presumed to be a mutually beneficial employment contract, replete with many other rules. It is a contract provision that helps both firms that adopt it and their workers who

[12]The analysis can really get sticky, and convoluted, when it is recognized that *commitments* that firms make are only implicitly made, with no formal contract, often with a host of unstated contingencies. For example, many firms may commit to overpaying their workers *if* the firm is not sold and *if* market conditions do not turn against them. Workers will simply have to consider those contingencies in the wages that they demand early in their careers and later on. All we can say is, the greater the variety and number of contingencies, the less the underpayment workers will accept early in their careers, and the less benefits firms and their workers will achieve from twisting the wage structure.

must abide by it. Parts of the contract can make the mandatory retirement rule economically sound.

And we have spent much of this section exploring the logic of twisting workers' career income paths. If such a twist is productive and profitable, and if workers must be overpaid late in their career to make the twist doable, then it follows that firms will want, at some point, to cut the overpayments off. What is mandatory retirement? *It is a means of cutting off at some definite point the stream of overpayments*. It is a means of making it possible, and economically practical, for a firm to engage a twisted pay scale and to improve incentives to add to the firm's productivity and profitability. To continue overpayments until workers -- even the most productive ones -- collapse on the job is nothing short of a policy that courts financial disaster.

Having said that, suppose Congress decides that mandatory retirement is simply an inane employment policy, as it has done? After all, members of Congress might reason, many of the workers who are forced to retire are still quite productive. What are the consequences?

Clearly, the older workers who are approaching the prior retirement age, who suffered through years of underpayment early in their careers, but who are, at the time of the abolition of mandatory retirement policy, being overpaid will gain from the passage of the law. They can continue to collect their overpayments until they drop dead or decide that work is something they would prefer not to do. They gain more in overpayments than they could have anticipated (and they get more back from their firms than they paid for in terms of their early underpayments). These employees will, because of the actions of Congress, experience an unexpected wealth gain.

There are, however, clear losers. The owners will suffer a wealth loss; they will have to continue with the overpayments. Knowing that, the owners will likely try to minimize their losses. Assuming that the owners can't lower their older workers' wages to market levels, and eliminate the overpayment (because of laws against age discrimination), the owners will simply seek to capitalize the expected stream of losses from keeping the older workers on and buy them out, that is, pay them some lump-sum amount to induce them to retire.

To buy the workers out, the owners would not have to pay their workers an amount equal to the current value of the workers' expected future wages. The reason is that the worker should be able to collect some lower wage in some other job if he or she is bought out. Presumably, the buyout payments would be no less than the value of the expected stream of *overpayments* (the pay received from the company minus the pay the worker could get elsewhere, appropriately discounted).

In order for the buyout to work, of course, both the owners and workers must be no worse off and, preferably, should gain by any deal that is struck. How can that be? Owners and workers could easily make a deal whereby both sides are no worse off. The owners simply pay the workers the current value of the overpayments (adjusted for the timing and uncertainty of the future payments).

But, can both sides *gain* by a buyout deal? That may not always be so easy to do. The owners would have to be willing to pay workers more than they, the workers, are willing to accept. There are several reasons such a deal may be possible in many, but not necessarily all, cases. First, the workers could have a higher discount rate than the owners, and this may often be the case because the owners are more diversified than their workers in their investments. Workers tend to concentrate their capital, a main component of which is *human capital*, in their jobs. By agreeing to a buyout and receiving some form of lump-sum payment in cash (or even in a stream of future cash payments), the workers can diversify their portfolios by scattering the cash among a variety of real and financial assets. Hence, workers might accept less than the current (discounted) value of their overpayments just to gain the greater security of a more diversified investment portfolio. Naturally (and we use that word advisedly), the workers cannot be sure how long they will be around to collect the overpayments. By taking the payments in lump-sum form, they reduce the risk of collection and increase the security of their heirs.

Second, sometimes retirement systems are overfunded, that is, they have greater expected income streams from their investments than are needed for meeting the expected future outflow of retirement payments. This is true, for example, of the California State Employee Retirement System. Therefore, if the company can tap the retirement funds, as the State of California did in the mid-1990s, it can pay workers more in the buyout than they would receive in overpayments by continuing to work. In so doing, they can move those salaries "off budget," which is what California has done in order to match its budgeted expenditures with declining funding levels for higher education.

Third, some workers may take the buyout because they expect their companies will meet with financial difficulty down the road of competition. The higher the probability the company will fail in the future (especially the near future), the more likely workers would be willing to accept a buyout that is less than the current value of the stream of overpayments

Fourth, some workers might take the buyout simply because they have tired of working for the company or want to walk away from built-up hostilities. To that extent, the buyout can be less than the (discounted) value of the overpayments.

Fifth, of course, older workers have to fear that the employer will not continue to pay workers more than they are worth indefinitely. The workers' fears arise from a combination of two factors: The owners can shuck their overpayments with a buyout. Then, the owners still maintain a great deal of discretion, in spite of any law that abolishes mandatory retirement rules. The owners can, if they choose to do so, lower the amount of the buyout payment simply by making life more difficult for older workers in ways that are not necessarily subject to legal challenge (for example, by changing work and office assignments, secretarial assistance, discretionary budgeted items, flexibility in scheduling, etc.).[13] The owners may never actually have to take such actions to lower the buyout payments. All that is necessary is for the *threat* to be a real consideration. Workers might rightfully expect that the greater their projected overpayments, the more they must fear their owners will use their remaining discretion to make a buyout doable.

We should also expect that workers' fears will vary across firms and will be related to a host of factors, not the least of which will be the size of the firm. Workers who work for large firms may not be as fearful as workers for small firms, mainly because large firms are more likely to be sued for any retaliatory use of their discretionary employment practices (and efforts to adjust the work of older workers in response to any law that abolishes mandatory retirement rules). Large firms simply have more to take as a penalty for what are judged to be illegal acts. Moreover, it appears that juries are far more likely to impose much larger penalties on large firms, with lots of equity, than their smaller counterparts. This unequal treatment before the courts, however, suggests that laws that abolish mandatory retirement rules will give small firms a competitive advantage over their larger market rivals.

However, we hasten to stress that all we have done is to discuss the transitory adjustments firms will make with their older workers, who are near the previous retirement age. We should expect other adjustments for younger workers, not the least of which will be a change in their wage structures. Not being able to overpay their older workers in their later years will probably mean that the owners will have to raise the pay of their younger workers. After all, the only reason the younger workers would accept underpayment for years is the prospect of overpayments later on.

There are three general observations from this line of inquiry that are interesting:

1. The abolition of mandatory retirement will tend to help those who are about to retire.
2. Abolition might help some older workers who are years from retirement, who work for large firms, and who can hang on to their overpayments. It can hurt other older workers who are fired, demoted, not given raises, or have their pay actually cut.

[13]Workers also understand that challenging the actions of owners can get expensive, which means that owners might take actions with regard to their older workers that are subject to legal challenge but only in a probabilistic sense. That is to say, owners might simply demote older workers. Even though employers who take such an actions *could* be taken to court, they might not be taken to court, given the expense the worker might have to incur and the likelihood that the challenge just might not be successful.

3. It can increase the wages of younger workers by lowering the amount by which they will be underpaid. However, their increase in wages while they are young will come at the expense of smaller overpayments later in their careers. Many, if not all, of these younger workers will not be any better off because of the abolition of mandatory retirement than they would have been with a retirement rule permitted.

Overall, productivity might be expected to suffer, given that owners can no longer twist their career pay structures for their workers. As a consequence, workers will not have as strong an incentive to improve their productivity. They simply cannot gain as much by doing so. This means that the abolition of mandatory retirement rules can lower worker wages from what they otherwise would have been.

The simple point that emerges from this line of discussion is that the level and structure of pay counts for reasons that are not always so obvious. But our point about "overpayment" is fairly general, applying to the purchase of any number of resources other than labor. You may simply want to "overpay" suppliers at times just to ensure that they will provide the agreed-upon level of quality, so that they will not take opportunities to shirk because they can lose, on balance, if they do so.[14]

The moral of the analysis is that most firms have good economic reasons for doing what they do. There are certainly solid economic grounds for overpaying workers, just as there are good reasons for mandatory retirement. We like to think that members of Congress were well intended when they abolished mandatory retirement rules back in 1978. Unfortunately, they simply did not think through these complex matters very carefully. (Perhaps the politics of the moment did not allow them to do so.) If they had considered the full complexity of firms' retirement policies, many older workers would not now be suffering through the impaired earnings and employment opportunities that members of Congress are now decrying.

Concluding Comments

The market is a system that provides producers with incentives to deliver goods and services to others. To respond to those incentives, producers must meet the needs of society. They must compete with other producers to deliver their goods and services in the most cost-effective manner.

A market implies that sellers and buyers can freely respond to incentives and that they have options and can choose among them. It does not mean, however, that behavior is totally unconstrained or that producers can choose from unlimited options. What a competitor can do may be severely limited by what rival firms are willing to do.

The market system is not perfect. Producers may have difficulty acquiring enough information to make reliable production decisions. People take time to respond to incentives,

[14] For a fuller discussion of how above-minimal price can give suppliers an incentive to provide above-minimal quality of products, see Benjamin Klein and Keith B. Leffler, "The Role of Market Forces in Assuring Contractual Performance," Journal of Law and Economics, vol. 89 (no. 4, 1981), pp. 615-641.

and producers can make high profits while others are gathering their resources to respond to an opportunity. In the electronics industry, three or four years were required to reduce the price of a basic calculator from $300 to $40. Some consumers still may not be getting exactly the kind of calculator they want.

An uncontrolled market system also carries with it the very real prospect that one firm will acquire monopoly power, restricting the ability of others to respond to incentives, produce more, and push prices and profits down.

In this chapter, we have paid a great deal of attention to how markets clear through a price set at the intersection of supply and demand. However, we have also noted that firms must be mindful of incentives in their methods of compensation. More specifically, we have indicated that, at times and under certain conditions, firms would be well advise not at every moment in time to match up worker pay with what workers are "worth." Current and prospective pay can be used as a means of increasing worker productivity and rewards over time. Similarly, mandatory retirement can also have unheralded benefits for workers as well as their employers. Mandatory retirement can allow for "overpayments" for workers, which can increase workers' incentives to improve their productivity over the course of their careers.

Review Questions

1. What are the consequences of competition in markets?
2. Why does the demand curve have a negative slope and the supply curve a positive slope?
3. "We know that markets don't always clear in the sense that the quantity supplied and demanded do not always match. Lines can be observed everywhere. Store shelves are often emptied or overstocked. Hence, why pay so much attention to the intersection of supply and demand?" Your task is to answer that question.
4. The mercantilists argued that a country's wealth consisted of its holdings of "gold bullion" (money). To keep gold in a country, they proposed tariffs and quotas to restrict imported goods and services.

   How do you react to that argument?
5. In what sense can competition in the production of undesirable goods be bad?
6. Why will the competitive market tend to move toward the price-quantity combination at the intersection of the supply and demand curves? What might keep the market from moving all the way to that equilibrium point?
7. Suppose you work for Levi Straus and the demand for blue jeans suddenly increases. Discuss possible short-run and long run movements of the market and the consequences for your company.
8. If the government imposes a price ceiling on gasoline, what would be the result? If

the price at the pump remains constant at the price ceiling, does that mean that the "real price" of gasoline has remained constant?

9. If the government imposes a price floor on whole milk and buys the resulting surplus, can it later sell what it has bought and recoup its expenditure? What else can the government do with the milk surplus? Why would you, as a milk producer, want the price floor? Show the industry benefits in a graph.
10. Henry Ford more than doubled his workers' wages. Did worker real income double by Ford's pay policy? Reflecting on the general principles behind Ford's pay action, when should any firm – your firm – stop raising the pay of workers (not in terms of actual dollar amount but in terms of some economic/management principle that you can devise)?
11. Workers and their employers often talk about how workers "earn" their wages but about how firms "give" their workers health insurance (or any other fringe benefit). Should the different methods of pay be discussed in different terms?
12. In state universities, why does the state subsidize full-time MBA programs but not executive MBA programs? Should the two programs be treated differently? Does the state subsidy explain the price differential for students in the two programs?

READING: **The Effect of Airline Deregulation on Travel Safety**

William F. Shughart II, University of Mississippi

Before 1978 airlines in the United States were strictly controlled by government agencies. The safety of airlines was, and remains today, regulated by the Federal Aviation Administration (FAA). In addition, the Civil Aeronautics Board (CAB) controlled airline fares and routes. The effect of CAB regulations was to restrict the ability of airlines to compete by price and entry into markets. Without CAB approval, for example, Delta Airlines could not lower its air fares or enter new markets to expand its business.

In 1978 Congress passed legislation to eliminate gradually most of the economic controls the CAB had over the domestic airline industry. However, airlines were not totally free to set prices and change routes until 1983.

Many commentators fear that airline deregulation may have resulted in a reduction in the safety of air travel in the United States.[1] From the perspective of economic theory, there are several reasons for believing that air safety may have been compromised. First, airline deregulation has led to reductions in the prices of many popular flights, especially long-distance flights (say, between New York and Los Angeles), and travel by air may have increased. Deregulation may have increased the opportunity for air accidents. Second, with the expansion of air travel, airlines may have had to draw on less experienced, qualified, and careful pilots and mechanics.

Third, with greater competition in the airline industry, several airlines may have become unprofitable and mangers may have reduced expenditures on needed plane repairs in order to increase airline profits. Fourth, before airfares were deregulated, airlines may have competed in many nonprice ways—for example, meals and in-flight service, movies, interiors of planes, and safety records. When they could compete by price after deregulation, airlines may have sacrificed safety competition for price competition. All of these factors may have led to increased air accidents and deaths.

Economists have statistically investigated the effect of airline deregulation on airline safety. While the debate continues, recent studies show that airline deregulation has in fact led to significantly more air travel but that the number of airline accidents and deaths has not been affected.[2] Airline deaths have been on a downward trend for decades, and airline deregulation does not appear (to date) to have slowed the pace of decrease.[3] Economists have reasoned that the greater freedom given airlines by deregulation may have been held in check by the considerable costs that airlines incur when they do have accidents. Airlines, in other words, may have continued to maintain their safety records because of the fear and cost of liability suits that are brought against them when they do have crashes. In addition, Congress never deregulated safety.

Various government policies often have hidden, secondary market effects that economists and policymakers must consider. Airline deregulation is a good case in point. Airline deregulation could have reduced total travel deaths in the country by its indirect impact on highway travel and accidents.

By deregulating airlines fares, Congress increased air travel. At the same time, Congress increased the *relative* cost of travel by car on the nation's highways. This is because, as noted, after deregulation, air travel became more convenient and often cheaper. Therefore, car travel became relatively expensive relative to air travel.

Airline deregulation has had two distinct effects on automobile travel. It has had a *price* (*or substitution) effect*. Less automobile travel would be expected with relatively lower airfares. Airline deregulation has also had an *income effect* because greater efficiency in air travel may have led to more national production and income. The greater national income may have led to more travel by air and cars.

Because the price and income effects of airline deregulation on automobile travel are not expected to be in the same direction, theory alone does not give a clear answer to the question, "How has airline deregulation affected automobile travel?" Statistical analysis is required, and the only study currently available on the issue found that airline deregulation has, indeed, reduced travel by automobiles (by an annual average of nearly 4 percent between 1979 and 1985).[4] However, because miles traveled on highways and automotive accidents and deaths are likely to be directly related, the small estimated decrease in automobile travel may have reduced automotive accidents and deaths by a sizable number. In fact, one of the authors estimates that airline deregulation has probably reduced automobile accidents by an annual average of several hundred thousand and deaths by an annual average of several hundred.[5]

The indirect effects of policy changes, which are revealed through economic analysis, cannot be ignored by policymakers. Policymakers need to be mindful of the fact that efforts to resurrect the type of airline regulation abandoned in the late 1970s may, or may not, improve airline safety records. Re-regulation, however, may cause people to shift from air travel to highway travel. Unfortunately, highway travel remains far more dangerous than air travel, and unless precautions are taken, overall travel deaths can be increased by airline re-regulation. This does not mean that re-regulation should not be undertaken but only that care must be taken in designing any new economic controls on airlines.

[1] See Hobart Rowen, "Bring Back Regulation," *Washington Post* (National Weekly Edition), August 31, 1987, p. 5.

[2] See Nancy L. Rose, *Financial Influences on Airline Safety*, no. 1890-87 (Cambridge, Mass.: Sloan School of Management, Massachusetts Institute of Technology, 1987; and Richard B. McKenzie and William F. Shughart II, "The Impact of Airline Regulation on Air Safety," *Regulation* (January 1988), pp. 42-47.

[3] Establishing the effect of airline deregulation [on] air travel and air accidents and deaths is more difficult than it appears. This is because many factors affect air travel and deaths, including the amount of income people in the economy have to spend. The very valuable statistical methods used by economists to separate the impact of airline deregulation [from] people's income are called econometrics.

[4] Richard B. McKenzie and John T. Warner, *The Impact of Airline Deregulation on Highway Safety* (St. Louis: Center for the Study of American Business, Washington University, December 1987).

[5] Ibid., p. 4.

CHAPTER 3

Principles of Rational Behavior at Work in Society and Business

We are not ready to suspect any person of being defective in selfishness.

Adam Smith

With this chapter we begin a detailed examination of key issues in microeconomics, namely the study of how prices are determined in individual markets. Prices are important – or, rather, should be important – to managers because of their unavoidable impact on the decisions of managers within individual firms. We have already seen how the forces of supply and demand determine prices (Chapter 2). Now we will explore the *determinants* of the supply and demand for goods, services, and resources.

Microeconomics rests on certain assumptions about individual behavior. One is that people are capable of envisioning various ways of improving their position in life. This chapter reviews and extends the discussion begun in Chapter 1 of how people – business people included -- go about choosing among those alternatives. According to microeconomic theory, consumers and producers make choices rationally, so as to maximize their own welfare and their firms' profits. This seemingly innocuous basic premise about human behavior will allow us to deduce an amazing variety of implications for business and every other area of human endeavor.

Rationality: A Basis for Exploring Human Behavior

People's wants are ever expanding. We can never satisfy all our wants because we will always conceive of new ones. The best we can do is to maximize our satisfaction, or utility, in the face of scarcity. **Utility** is the satisfaction a person receives from the consumption of a good or service or from participation in an activity. Happiness, joy, contentment, or pleasure might all be substituted for satisfaction in the definition of utility. Economists attempt to capture in one word—utility—the many contributions made to our well being when we wear, drink, eat, or play something.

The ultimate assumption behind this theory is that people act with a purpose. In the words of von Mises, they act because they are "dissatisfied with the state of affairs as it prevails."[1]

---

[1] Ludwig von Mises, *The Ultimate Foundations of Economic Science: An Essay on Method* (Princeton, N.J.: D. Van Nostrad, 1962), pp. 2—3.

The Acting Individual

If people act in order to satisfy their consciously perceived wants, their behavior must be self -- directed rather than externally controlled. However, there is no way to prove this assertion. Economists simply presume that individuals, as opposed to groups, perform actions. It is the individual who has wants and desires, and looks for the means to fulfill them. It is the individual who attempts to render his or her state "less unsatisfactory."

Group action, when it occurs, results from the actions of the individuals in the group. Social values, for instance, draw their meaning from the values held collectively by individuals. Economists would even say that group action cannot be distinguished from individual action. Although economists do not deny the existence of group psychology, they leave the study of social groups to others. Thus to understand group behavior, the economist looks to the individual.

Of course individuals in a group affect one another's behavior. In fact, the size and structure of a group can have a dramatic effect on individual behavior. When economists speak of a competitive market, they are actually talking about the influence that other competitors have on the individual consumer or firm.

Rational Behavior

When individuals act to satisfy their wants, they behave rationally. **Rational behavior** is consistent behavior that maximizes an individual's satisfaction. The notion of rational behavior rests on three assumptions:

- First the individual has a preference and can identify, within limits, what he or she wants.
- Second, the individual is capable of ordering his or her wants consistently, from most preferred to least preferred.
- Third, the individual will choose consistently from these ordered preferences to maximize his or her satisfaction.

Even though the individual cannot fully satisfy all her wants, she will always choose more of what she wants rather than less. Furthermore, she will always choose less rather than more of what she does not want. In short, the rational individual always stands ready to further her own interests.

Some readers will find these assertions obvious and acceptable. To others, they may seem narrow and uninspiring. Later in the chapter we will examine some possible objections to the concept of rational behavior, but first we must examine its logical consequences.

Rational Decisions in a Constrained Environment

Several important conclusions flow from the economist's presumption of rational behavior. First, the individual makes choices from an array of alternatives. Second, in making each choice, a person must forgo one or more things for something else. All rational behavior involves a cost, which is the value of the most preferred alternative forgone. Third, in striving to maximize his or her welfare, the individual will take those actions whose benefits exceed their costs.

Choice

We assume that the individual can evaluate the available alternatives and select the one that maximizes his utility. Nothing in the economic definition of rational behavior suggests that the individual is completely free to do as he wishes. Whenever we talk about individual choices, we are actually talking about constrained choices—choices that are limited by outside forces. For example, you as a student find yourself in a certain social and physical environment and have certain physical and mental abilities. These environmental and personal factors influence the options open to you. You may have neither the money, the time, nor the stomach to become a surgeon, or your career goal may not allow you the luxury of taking many of the electives listed in your college catalog.

Although your range of choices may not be wide, choices do exist. At this moment you could be doing any number of things instead of reading this book. You could be studying some other subject, or going out on a date, or playing with your son or daughter. You could have chosen to go shopping, to engage in intramural spots, or to jog around the block. You may not be capable of playing varsity sports, but you have other choices. Although your options are limited, or constrained—you are not completely free to do as you please—you can still choose what you want to do. In fact, you must choose.

Suppose that you have an exam tomorrow in economics and that there are exactly two things you can do within the next 12 hours. You can study economics, or you can play your favorite video game. These two options are represented in Figure 3.1. Suppose you spend the entire 12 hours studying economics. In our example, the most you can study is four chapters, or E_1. At the other extreme, you could do nothing but play games—but again, there is a limit: eight games or G_1.

Neither extreme is likely to be acceptable. Assuming that you aim both to pass your exam and to have fun, what combination of games and study should you choose? The available options are represented by the straight line E_1G_1, the production possibilities curve for study and play and the area underneath it. If you want to maximize your production, you will choose some point on E_1G_1, such as a: two chapters of economics and four games. You might yearn for five games and the same amount of study, but that point is above the curve and beyond your capabilities. If you settle for less—say one chapter and three games, or point x—you will be doing less than you are capable of doing and will not be maximizing your utility. The combination you actually choose will depend on your preference.

Changes in your environment or your physical capabilities can affect your opportunities and consequently the choices you make. For example, if you improve your study skills, your production rate for chapters studied will rise. You might then be able to study eight units of economics in 12 hours -- in which case your production possibilities curve would expand outward. Even if your ability to play Amazons from Outer space remained the same, your greater proficiency in studying would enable you to increase the number of games played. Your new set of production possibilities would be E_2G_1 in Figure 3.2.

Again, you can choose any point along this curve or in the area below it. You may decide against further games and opt instead for four chapters of economics (point *c*). You could move to point *b*, in which case you would still be learning more economics—three chapters instead of two—but would also be playing more games. The important point is that you are able to choose from a range of opportunities. The option you take is not predetermined.

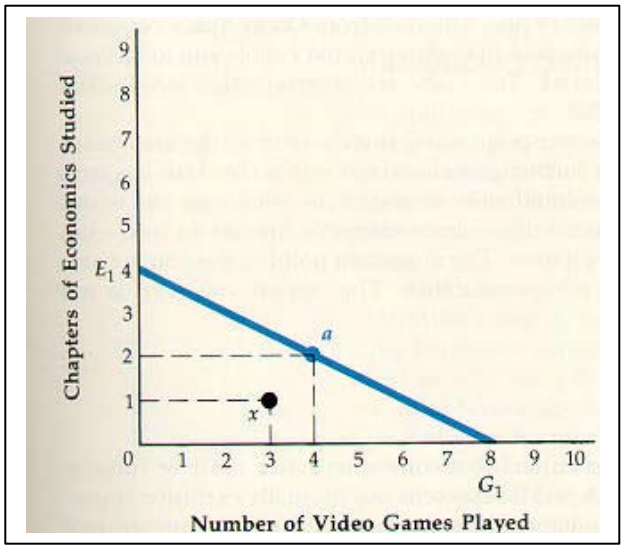


FIGURE 3.1 Constrained Choice

With a given amount of time and other resources, you can produce any combination of study and games along the curve E_1G_1. The particular combination you choose will depend on your personal preferences for those two goods. You will not choose point *x*, because it represents less than you are capable of achieving—and as a rational person, you will strive to maximize your utility. Because of constraints on your time and resources, you cannot achieve a point above E_1G_1.

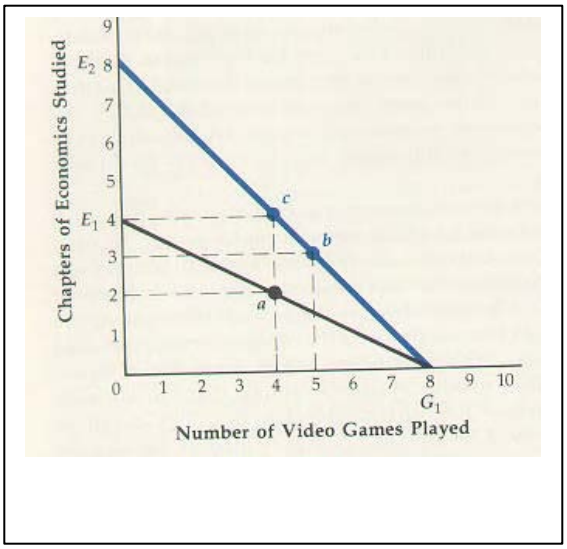


FIGURE 3.2 Change in Constraints

If your study skills improve and your ability at the game remains constant, your production possibilities curve will shift from E_1G_1 to E_2G_1. Both the number of chapters you can study and the number of games you can play will increase. On your old curve, E_1G_1, you could study two chapters and play four games (point *a*). On your new curve E_2G_1, you can study three chapters and play five games (point *b*).

Cost

The fact that choices exist implies that some alternative must be forgone when another is taken. If A and B represent two mutually exclusive opportunities, to choose A is simultaneously to not choose B. In the presence of choice—a situation in which no more than one alternative can be taken at a time—a cost must be incurred. **Cost** (or more precisely, opportunity cost) is the value of the most highly preferred alternative not taken. Put another way, it is the value the individual places on the most favored alternative not taken at the time the choice is made. For example, suppose that you have decided to spend half an hour watching old television programs. The two programs you most want to watch are *M.A.S.H.* and *Gilligan's Island*. If you choose *Gilligan's Island*, the cost is the pleasure you sacrifice by not watching *M.A.S.H.*

Notice that cost is not defined in terms of money. Money is a useful measure because it reduces all costs to one common denominator. Money is only the means of measuring cost, however; it is not cost itself. The shoes you are wearing may have cost you $50 (a money cost), but the real cost (the opportunity cost) is the value of what you could have purchased instead. **Money cost** is a monetary measure of the benefits forgone when a choice is made. The real cost is the actual benefits given up from the most preferred alternative not taken when a choice is made. When economists use the term cost, they mean real, or opportunity, cost. You could have bought dozens of soft drinks or deposited the $50 in a savings account for future use. Either option would be a legitimate alternative to purchasing shoes. The point is that the cost of the shoes to you is the value of the most attractive option not taken, whether it is the soft drinks or the future use of the money.

As long as you have alternative uses for your time and other resources, there is no such thing as a free lunch. Nothing can be free if other opportunities are available. One goal of economics courses is to help you recognize this very simple principle and to train you to search for hidden costs. There is a cost to writing a poem, to watching a sunset, to extending a common courtesy, if only to open a door for someone. Although money is not always involved in choices, the opportunity to do to other things is. A cost is incurred in every choice.

Maximizing Satisfaction: Cost-benefit Analysis

An individual who behaves rationally will choose an option only when its benefits are greater than or equal to its costs. Furthermore, individuals will try to maximize their satisfaction by choosing the most favorable option available. That is, they will produce or consume those goods and services whose benefits exceed the benefits of the most favored opportunity not taken.

This restatement of the maximizing principle, as it is called, explains individual choice in terms of cost. In Figure 3.1, the choices along curve E_1G_1 represent various

cost-benefit tradeoffs. If you choose point *a*, we must assume that you prefer *a* to any other combination because it yields the most favorable ratio of benefits to costs.

A change in cost will produce a change in behavior. Suppose you and a friend set a date to play checkers, but at the last moment he received a lucrative job offer for the day of the match. Most likely the contest will be rescheduled. The job offer will change your friend's opportunities in such a way that what otherwise would have been a rational act (playing checkers) becomes one that is no longer rational. The cost of playing checkers will rise significantly, enough to exceed the benefits of most checkers games.

Economists see cost-benefit analysis as the basis of much (but certainly not all) of our behavior. **Cost-benefit analysis** is the careful calculation of all costs and benefits associated with a given course of action. Why do you attend classes, for example? The obvious answer is that at the time you decide to attend class, you expect the benefits to attending the exceed the costs. The principle applies even to classes you dislike. A particular course may have no intrinsic value, but you may fear that by cutting class, you will miss information that would be useful on the examination. Thus the benefits of attending are a higher grade than you would otherwise expect. Besides, other options open to you on Tuesday morning at 10:00 AM may have so little appeal that the cost of going to class is very slight.

Take another example. Americans are known for the amount of waste they pile up. Our gross national garbage is estimated to be more valuable than the gross national output of many other nations. We throw away many things that people in other parts of the world would be glad to have. However morally reprehensible, waste may be seen as the result of economically rational behavior. Wastefulness may be beneficial in a limited personal sense. The food wrappings people throw away are "wasted," but they do add convenience and freshness to the food. In the individual's narrow cost-benefit analysis, the benefits of the wrapping can exceed the costs.

Is life priceless? Although we like to think so, many of us are not willing to bear the cost that must be paid to preserve it. Several million animals—dogs, opossums, squirrels, and birds—are killed on the highways each year. Most of us make some effort to avoid animal highway deaths. If saving lives were all -- important, we could drive less -- but that would bring a significant cost. Even when human beings are involved, we sometimes refuse to bear the cost of preserving life. People avoid helping victims of violent crime, and doctors routinely pass by highway accidents although they might save lives by stopping to help. Indeed, revolutions succeed through people's willingness to sacrifice lives—both others' and their one -- to achieve political or economic goals.

The behavior of business people is not materially different from that of drivers or consumers. People in business are constantly concerned with cost-benefit calculations, only the comparisons are often (but not always) made in dollar terms: For example, whether the cost of improving the quality of a product is matched by the benefits of the improvement. Will consumers value the added benefits enough to pay for hem? In assessing the safety of their products, business people must consider whether consumers are willing to pay the cost of any improvements.

The Effects of Time and Risk on Costs and Benefits

When an individual acts, costs are not necessarily incurred immediately, and benefits are not necessarily received immediately. The decision to have a child is a good example. At turn of the century prices, a college -- educated couple's first child can easily cost more than $500,000, from birth through college.[2] Fortunately this high cost is incurred over a relatively long period of time (or people would rarely become parents!).

Benefits received in the future must also be compared with present benefits. If you had a choice between receiving $10,000 now and $10,000 one year from now, you would take $10,000 today. You could put the money in a bank, if nothing else, where it would earn interest, or you could avoid the effects of future inflation by spending the money now. In other words, future benefits must be greater than present benefits to be more attractive than present benefits.

To compare future costs and benefits on an equal footing with costs and benefits realized today, we must adjust them to their present value. **Present value** is the value of future costs and benefits in terms of current dollars. The usual procedure for calculating present value -- a process called discounting -- involves an adjustment for the interest that could be earned (or would have to be paid) if the money were received (or due) today rather than in the future.[3]

If there is any uncertainty about whether future benefits or costs will actually be received or paid, further adjustments must be made. Without such adjustments, perfectly rational act may appear to be quite irrational. For example, not all business ventures can be expected to succeed. Some will be less profitable than expected or may collapse altogether. The *average* fast-food franchise may earn a yearly profit of $1 million, but, but only nine out of ten franchises may survive their first year (because the *average* profits is distorted by the considerable earnings of one franchise). Thus the estimated profits for such a franchise must be discounted, or multiplied by 0.90. If 10 percent of such ventures can be expected to fail, on average each will earn $900,000 ($1 million x .90).

The entrepreneur who starts a single business venture runs the risk that it may be the one out of ten that fails. In that case profit will be zero. To avoid putting all their eggs in one basket, many entrepreneurs prefer to avoid putting all their "eggs" in the

---

[2] For rough estimates of the cost of rearing children by expenditure, see U.S. Department of Commerce, Statistical Abstract of the United States: 1998 (Washington, D.C.: U.S. Government Printing Office, 1998), table 732. To obtain the total cost of childcare, you must then estimate the value of parental time.

[3] The mathematical formula for computing the present value of future costs or benefits received one year from now is PV = [1/(1 + *r*)] *f*, where PV stands for present value, *r* for the rater of interest, and *f* for future costs or benefits. The interest rate used in this formula is the rate at which we discount future costs and benefits.

proverbial one basket by initiating several new ventures, thereby spreading the risk of doing business. In the same way, investors spread their risk by investing in a wide variety of companies, and firms spread their risk by producing a number of products.

To give another example, criminal behavior may appear irrational if only the raw costs and benefits are considered. A burglar who nets $1,500 from the sale of stolen property may have to spend a year in jail if caught, prosecuted, and convicted. He could lose the annual income from his legitimate job, perhaps $10,000. That is a high cost to pay for a $1,500 profit on stolen property, but he pays that cost only if he is caught, prosecuted, convicted, and sentenced. The police cannot be everywhere at all times; prosecutors may be reluctant to prosecute; and suspended sentences are commonplace. All in all, even an inept burglar may have no more than a 10 percent chance of spending a year in jail.[4]

To estimate the actual cost faced by the burglar who is caught, sentenced, and sent to jail for a year, we might multiply the cost if caught, $10,000, by 0.10. That calculation indicates that to a burglar who is sent to jail for an average of one out of ten burglaries, the cost of any one burglary is only $1,000 ($10,000 x 0.10). Thus the actual cost of the burglary is less than the benefits received, $1,500. Although it may be morally reprehensible, the criminal act can conceivably be a rational one.

Surveys of criminal activities and their rewards tend to support such a conclusion. A study of burglary and grand larceny cases in Norfolk, Virginia, showed that for the unusual criminal who committed just one crime and was caught in the act, crime did not pay. The typical criminal, however, convicted the average number of times and sentenced to the average number of years in prison, more than tripled the lifetime income he could have earned from a regular salaried job—even allowing for one or more years of unsalaried incarceration.[5] When this study was replicated in Minnesota, the results were not quite as dramatic, but the criminal's lifetime income still doubled.[6] For criminals who are never caught, crime pays even more handsomely.

The same logical process of discounting can be applied to your life as a student. When you signed up for your MBA program, you actually had limited information on how it would work out for you. (Admit it, it was a gamble!) Similarly, when you sign up for courses, you usually have only a very rough idea of how difficult and time --

---

[4] This is not an unreasonably low figure. Gregory Krohm "concluded that the chance of an 'adult' (seventeen or older) burglar being sent to prison for any single offense is .0024. . . For juveniles. . . the risk was much lower, .0015." "The Pecuniary Incentives of Property Crime," in *The Economics of Crime and Punishment*, ed. Simon Rottenberg (Washington, D.C.: American Enterprise Institute for Public Research, 1973), p. 33.

[5] William E. Cobb, "Theft and the Two Hypotheses," in The Economics of Crime and Punishment, ed. Simon Rottenberg (Washington, D.C.: American Enterprise Institute for Public Policy Research, 1973), pp. 19 -- 30.

[6] David L. Johnson, "An Analysis of the Costs and Benefits for Criminals in Theft" (Economics Department, St. Cloud State College, St. Cloud, Minn., May 1974), mimeographed.

consuming they will be, and what benefits you will receive from them. In other words, you are rarely certain of their costs and benefits. To make your decision, you will have to discount the raw costs and benefits by the probability of their being realized. Risks are pervasive in human experience, and rational behavior takes those risks into account.

What Rational Behavior Does *Not* Mean

The concept of rational behavior often proves bothersome to the noneconomist. Most of the difficulties surrounding this concept arise from a misunderstanding of what rationality means. Common objections include the following:

1. People do many things that do not work out to their benefit. A driver speeds and ends up in the hospital. A student cheats, gets caught, and is expelled from school. Many other examples can be cited. To say that people behave rationally does not mean that they never make mistakes. We can calculate our options with some probability, but we do not have perfect knowledge, nor can we fully control the future. Chances are that we will make a mistake at some point, but as individuals, we base our choices on what we expect to happen, not on what does happen. We speed because we expect not to crash, and we cheat because we expect not to be caught. Both can be rational behaviors.

2. Rational behavior implies that a person is totally self -- centered, doing only things that are of direct personal benefit. Rational behavior need not be selfish. Altruism can be rational; a person can want to be of service to others, just as he can want to own a new car. Most of us get pleasure from seeing others happy—and particularly when their happiness is the result of our actions. Altruism may not always spring from rational cost-benefit calculations; however, it is not always inconsistent with economic rationality. Self -- interest, moreover, does not necessarily stop at the individual. For many actions, "self" includes members of one's family or friends. When a father spends a weekend building a tree house for his children, economists say that he has been engaged in self -- interested behavior.

3. People's behavior is subject to psychological quirks, hang -- ups, habits and impulses. Surely such behavior cannot be considered rational. Human actions are governed by the constraints of our physical and mental makeup. Like our intelligence, our inclination toward aberrant or impulsive behavior is one of those constraints. It makes our decision-making less precise and contributes to our mistakes, but it does not prevent our acting rationally. Moreover, what looks like impulsive or habitual behavior may actually be the product of some prior rational choice. The human mind can handle only so much information and make only so many decisions in one day. Consequently, we may attempt to economize on decision making by reducing some behaviors to habit. Smoking may appear to be totally impulsive, and the physical addition that accompanies it may indeed restrict the smoker's range of choices. Why might a person pull a cigarette from the pack "without

thinking"? Perhaps because she has reasoned earlier that contemplating the pros and cons of smoking each and every time she things of cigarettes is too costly. By allowing smoking to become more or less automatic, the smoker probably increases the number of cigarettes she smokes daily, but she sees the tedium of having to make the decision each and every time she smokes.

4. Rational behavior implies that people know what they want, that they know which alternatives are available, and that they know how to act on that information. People cannot assimilate all the information they need to make rational choices, however. People do lack information, and they could make better choices if information were easier to obtain. However, rational behavior does not require perfect information. People will make choices on the basis of the information they have or can rationally acquire. If they have less than perfect information, they may make mistakes in their choices. The success or failure of their choices must be judged within those constraints.

5. People do not necessarily maximize their satisfaction. For instance, many people do not perform to the limit of their abilities. Satisfaction is a question of personal taste. To some individuals, lounging around is an economic good; by consuming it, they increase their welfare. Criticism of such is tinged with normative value judgments. An observer who equates rational behavior with what he or she considers good will have no trouble demonstrating that such behavior is irrational. **Irrational behavior** is behavior that is inconsistent or clearly not in the individual's best interests and that the individual recognizes as such at the time of the behavior.

6. But to the economist, the values of the actor, not the observer or the social critic, determine the rationality of an act. Harold, not Jennifer or Max, determines the rationality of Harold's behavior.

Disincentives in Poverty Relief

Our discussion of rational behavior can be used to understand one of the biggest policy issues of our time, welfare reform. We can do this by assuming that welfare recipients are tolerably rational.

So much of the public discussions about welfare programs, especially cuts in them, assumes that since Congress has the authority to change the programs, it can alter the programs any way it wishes without creating problems. However, as we can easily see, Congress is in something of an economic, if not political, bind on welfare relief, given how incentives change when the program is adjusted. The basic problem is that the practice of scaling down welfare benefits as earned income rises creates an implicit marginal tax on additional earned income that discourages the poor from working. Why not lower the implicit marginal tax rate?

Figure 3.3 gives the answer. The 45 -- degree line that extends out from the origin indicates points of equal distance from each axis—that is, points at which spendable

income equals earned income. At point *y*, for example, a poor person earns and can spend $5,000 annually. At points above the line, spendable income exceeds earned income. For instance, at point *x*, a poor person earns $5,000 annually and can spend $7,500. He receives a subsidy equal to $y - x$, or $2,500.

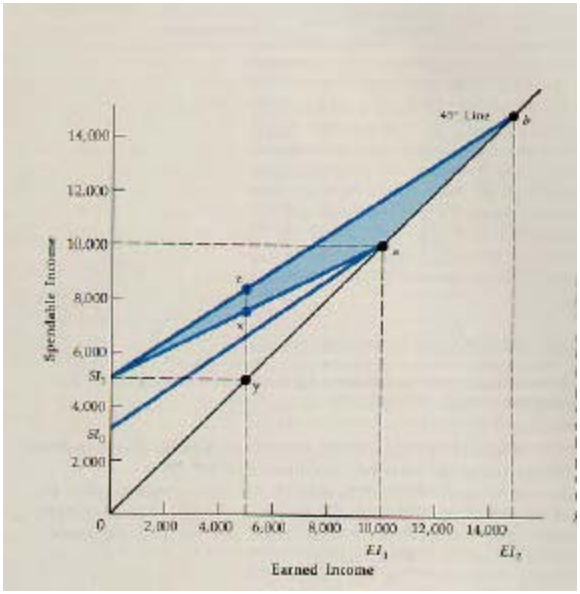


FIGURE 3.3 Policy Tradeoffs of a Negative Income Tax

With a guaranteed income of SI_1 ($5,000) and a break -- even earned income level of EI_1 ($10,000), the implicit marginal tax rate on the poor is 50 percent. If policymakers attempt to reduce the implicit tax rate by raising the break -- even income level, however, the government's poverty relief budget will rise by the shaded area $SI_1 ab$. A higher explicit tax burden will fall on a smaller group of taxpaying workers.

Suppose the government establishes a negative income tax with a guaranteed annual income level of $5,000, or SI_1. The break -- even earned income level is $10,000, or EI_1. A person who earns nothing will receive a subsidy of $5,000 a year. As his earned income rises the subsidy will decline, until it reaches zero at $10,000. Curve SI_1 shows the spendable income of people in this program at various earned income levels. They lose $500 in subsidies for every $1,000 of additional earned income. That is, they face an implicit marginal tax rate of 50 percent.

If policy markers want to reduce the implicit marginal tax rate on an earned income of $10,000 to less than 50 percent, they must either reduce the guaranteed spendable income level or raise the break -- even earned income level. If they raise the break -- even earned income level—to $15,000, or EI_2, for example—curve $SI_1 a$ will shift to $SI_1 b$. But then more people—all those with earned incomes up to EI_2—will receive benefits. Moreover, all the people covered originally will receive larger subsidies. A person with an income of $5,000 would receive $8,000 instead of $7,000 in spendable income (point *z* instead of point *x*), for example. The total increase in the government's poverty relief expenditures would equal the shaded area in the figure bounded by $SI_1 ab$.

The increase in expenditures would place a greater tax burden on taxpaying workers. Yet because more workers would be covered by the negative income tax, fewer people

would share the increased tax burden. Thus the explicit marginal tax rate on high -- income workers rises—lowering their incentive to work and earn additional income.

If the government reduces the guaranteed income level, say from SI_1 to SI_0, a different problem will result. On the new curve SI_0 *a*, the poor will receive less government aid at each earned income level. They may have more incentive to work under such an arrangement, but will they have enough to live on?

Policymakers, then, face difficult tradeoffs between the goal of helping the poor and the goal of minimizing the disincentive to work. To provide adequate aid, they may have to raise the breakeven income level high enough that people who are not strictly poor benefit. Yet to reduce aid to people who are not truly poor, they would have to lower the break -- even income level—thus increasing the implicit tax rate on the poor. To keep the implicit marginal tax rate down, they could lower the guaranteed income level—decreasing the benefits that go to the truly poor.

Our graphic analysis suggests that there may be economic as well as altruistic limits to the government's ability to transfer income from the rich to the poor. As more and more income is allocated to the poor, either the guaranteed income or break -- even income level must go up. If only the guaranteed income level is raised, the implicit marginal tax rate facing the poor increases. If that problem is avoided by raising the break -- even income level, poverty relief will cover more people, and the taxes paid by the remaining workers will go up. Increased aid to the poor thus should have three consequences. A higher explicit tax burden will fall on fewer taxpayers. Because of this burden, higher -- income groups will have less incentive to work, and lower -- income groups, because of the higher implicit tax rate, will also be less inclined to work.

MANAGER'S CORNER: **The Last-Period Problem**

Much of this chapter has been concerned with how people behave rationally. Here, we introduce "opportunistic behavior" as a form of rational behavior that people in business will want to protect themselves from. We suggest ways different parties to business deals can take advantage of other parties and how managers can structure their organizational and pay policies to minimize what we call "opportunistic behavior." More specifically, this section is concerned with how an announced end to a business relationship can inspire opportunistic behavior. Its goal is, however, constructive, structuring business deals – and the embedded incentives -- in order to maximize the durability and profitability of the deals. To do that, business relationships must be ongoing, or have no fixed end, to the extent possible. Having a fixed termination date can encourage opportunistic behavior, which can reduce firm revenues and profits. That is to say, a reputation for continuing in business has economic value, which explains why managers work hard to create such a reputation.

Problems with the End of Contracts

A terrific *advantage* of dealing with outside suppliers is that the relationship is constantly up for renewal and can easily be terminated if it is not satisfactory to both parties. But therein lies an important *disadvantage* of dealing with outside suppliers: the relationship lacks permanence or confidence that any given buyer/supplier relationship will be renewed. The supplier must attribute some probability that the end of the contract will be the end of the relationship, given that he or she might not be the next low bidder, a deduction that can have profound effects on the relationship that the astute manager must recognize. Without much question, firms have begun to develop relationships with suppliers that approximate partnerships because of the "last-period" problems inherent in relationships that are totally grounded in the low -- bidder status of the suppliers.

The basic problem is that during the last period of any business relationship, there is no penalty for cheating, which implies maximum incentive to cheat. As a consequence, cheating on deals in the last period is more likely than at any other time in the relationship.

Consider a simple business deal. Suppose that you want a thousand widgets of a given quality delivered every month, starting with January and continuing through December, and that you have agreed to make a fixed payment to the supplier when the delivery is made. If you discover after you have made payment that your supplier sent fewer than a thousand units or sent the requisite thousand units but of inferior quality, you can simply withhold future checks until the supplier makes good on his or her end of the bargain. Indeed, you can terminate the yearlong contract, which can impose a substantial penalty for any cheating early in the contract. Knowing that, the supplier will tend to have a strong incentive early on in the contract period to do what he or she has agreed to do.

However, the supplier's incentive to uphold his or her end of the bargain begins to fade as the year unfolds, for the simple reason that there is less of a penalty -- in terms of what is lost from your ending the working relationship -- that you can impose. The supplier might go so far as to reason that during the last period (December), the penalty is very low, if not zero. The supplier can cut the quantity or quality of the widgets delivered during December and then can take the check before you know what has been done. The biggest fear the supplier has is that you might inspect the shipment before handing over the final check. You may be able to get the supplier to increase the quantity or quality somewhat with inspection, but you should expect him or her to be somewhat more difficult to deal with. And you should not *expect* the same level of performance or quality.

The problem is that you have lost a great deal of your bargaining power during that last month, and that is the source of what we call and mean by the **last-period** (or end -- period) **problem**, meaning the costs that can be expected to be incurred from opportunistic behavior when the end of a working relationship approaches. It is a problem, however, that can be mitigated in several ways. The simplest and perhaps most common way is by maintaining continuing relationships. If you constantly jump from one supplier to another, you might save a few bucks in terms of the quoted prices, but

you might also raise your costs in terms of unfulfilled promises by suppliers during the last period of their association with you. "Working relationships," in other words, have an economic value apart from what the relationship actually involves, for example, the delivery of so many widgets. This is one important reason businesses spend so much time cultivating and maintaining their relationships and why they may stick with suppliers and customers through temporary difficulties.

Solutions to the Last-Period Problem

Nothing works to solve the last-period problem, however, like success. The more successful a firm is -- the greater the rate of growth for the firm and its industry -- the more likely others will recognize that the firm will continue in business for sometime into the future. The opposite is also true -- failure can feed on itself as suppliers, buyers, and workers begin to think that the last period is near. Firms understand these facts of business life. As a consequence, executives tend to stress their successes and downplay their failures. Their intent may not be totally unethical, given how bad business news can cause the news to get worse. Outsiders understand these tendencies. As a consequence, many investors pay special attention to whether executives are buying or selling their stock in their companies. The executives may have access to (accurate) insider information that is not being distributed to the public.

Another simple way of dealing with the last-period problem in new relationships is to leave open the prospect of future business, in which case the potential penalty is elevated (in a probabilistic sense) in the mind of the supplier. When there is no prospect of future business, the *expected* cost from cheating is what can be lost during the last period. When there is some prospect of future business, the cost is greater, equal to the cost that can be imposed during the last period plus the cost (discounted by the probability that it will be incurred) incorporated in the loss of future business.

When dealing with remodeling or advertising firms, for instance, you can devise a contract for a specified period, but you can suggest, or intimate, in a variety of creative ways, that if the work is done as promised and there are no problems, you might extend the contract or expand the scope of the relationship. In the case of the remodeling firm, you might point out other repairs in the office that you are thinking of having done. In the case of the advertising firm, you might suggest that there are other ad campaigns for other products and services that you are considering.

You should, therefore, be able to secure somewhat better compliance with your supplier during the last period of the contract, and how much the compliance is improved can be related to just how well you can convince your supplier that you mean business (and a lot of it) for some time into the future. However, we are not suggesting that you should outright lie about uncertain future business. The problem with lying is that it can, when discovered, undercut the value of your suggestions of further business and bring back to life the last-period problem. You need, in other words, to be prepared to extend, from time to time (if not always), working relationships when in fact they work the way you want them to work.

However, if you are not able to develop that impression, the last period can come sooner than you might think (or sooner than December in our earlier example). That is, the contractual relationship can unravel because of the way you and the supplier begin to *think* about what the other is thinking and how the other might act as a consequence.

If both you and your supplier are inclined to cheat on the contract, and you have already figured that your supplier will cheat to the maximum (send nothing) during the last period, then December becomes irrelevant and November becomes the last period. Your incentive then is to cheat on the supplier in November. Well, with November now the last period, you can imagine what your supplier is thinking. He is contemplating cheating in November before you get a chance to cheat. Ah, but you can bet the supplier by cheating in October. That thought suggests that when contemplating the contract before it is signed and sealed, you and the supplier can reach the conclusion that January is the (relevant) last period -- which means that the deal will never be consummated. In this way, the *last-period* problem becomes a *first* -- period problem, actually one of setting the terms of the contract. This way of thinking about it can make the signing problematic, and more costly than it need be, assuming there are ways around the problem.

This line of argument reminds us of an old joke about a prisoner condemned to death. As it happened, the prisoner was told on Sunday that he would be hung between Monday and Saturday, *but the day of his hanging would be a total surprise*. He reasoned, "They can't hang me on Saturday because it wouldn't be a surprise. So, Friday is the last day of the relevant period." Therefore, he reasoned, "They can't hang me on Friday because if they wait until then, it won't be a surprise." Continuing this line of reasoning, Friday gave way to Thursday being the last day, and so forth. He eventually concluded that they couldn't hang him. Of course, when they hung him on Wednesday, *he was really surprised*!

This joke suggests that the last period problem doesn't always lead to an unraveling in which the last period becomes the first. But the last period problem is potentially serious and is one reason that firms exist: firms are collections of departments (and people) who have continuing relationships that are not always up for re -- bidding, which means that the parties can figure that they will be continued, with there being no clear last period. The last-period problem is also a significant reason why the *corporation* is such an important form of doing business. The corporation is a legal entity whose existence is independent of the life of the owner or owners; the corporation typically lives on beyond the death of the owners. Given that ownership is in shares, the corporation makes for relatively easy and seamless transfer of ownership, which means the life of the company is, in an expectational sense, longer as a corporation than as a partnership or proprietorship, two organizational forms that die with the owners. This means that the corporate charter should be prized simply because it adds value to the company by muting (though not always eliminating) the last-period problem.

The last-period problem extends beyond buyer -- supplier relationships of the sort we described above involving the purchase of widgets. There is clearly a last-period problem for military personnel. When officers or enlisted men and women are given

their transfer orders, they can sit back and relax, given that the penalties that can then be imposed on them have been severely limited by the orders to move on. The problem becomes especially severe when personnel are about to leave the military altogether. Military people have a favorite expression for what we call shirking during the last period. They call it "FIGMO": "F -- -- k you, I've got my orders." We are sure that the military has devised a variety of ways to mute the impact of FIGMO, but it is equally clear that the problem of shirking as military men and women approach the ends of their assignments remains a pressing one. Sometimes you just have to accept some costs of shirking (otherwise you might end up concluding that people should be fired the moment they enlist, which can be more costly than the shirking).

The last-period problem can surface with a vengeance when an employee who has access to easily destroyed records and equipment is fired. The firm doing the firing must worry that the employee will use his or her remaining time in the plant or office to impose costs on the firm, to "get back" at the firm. As a consequence, firings are often a surprise, done quickly, with the employee given little more time than to collect his or her personal things in the office – all to minimize damage. The firm may even hand the employee a paycheck for hours of work not done, simply to make the break as quickly as possible and discourage fired workers from imposing even greater costs through damage to records and equipment. Indeed, when the potential for serious damage is present and likely, firms may hire a security guard to be with the fired employee until he or she is escorted to the door for the last time.

The last-period problem can also show up in the greater incentives people have to shirk as they approach retirement. To prevent workers from shirking, deferred compensation can be used with some of the compensation withdrawn if shirking ever does occur. A variation of this type of solution for executives is to tie their compensation to stock. If executives shirk toward the ends of their careers, causing their companies to do poorly, then the executives lose more than any remaining salary they are due for the duration of their tenure; they lose the value of the stock, which approximated the discounted value of the company's lost earnings attributable to the executives' shirking while still on the job.

Apparently, corporations' executive compensation committees are aware of the last-period problem. Economists Robert Gibbons and Kevin Murphy have found from their econometric studies that as CEOs get closer to retirement age, their compensation tends to become more closely tied to their firm's stock market performance.[7]

Another way of solving the last-period problem is through performance payments, which means that payments are made as a project is completed. For example, separate payments can be made for constructing a house when the house is framed, when it is under roof, and when wiring is in and the interior walls have been finished. However, a significant portion of the total amount due is withheld until after the entire project is

[7] Robert Gibbons and Kevin J. Murphy, "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence," Journal of Political Economy, vol. 100, n3 (June 1992), pp. 468 -- 506.

completed and the results approved. For example, 20 percent of the entire construction cost is not paid until after the final inspection.

Business critics often decry the extent to which many pension plans are not fully "funded" -- that is, not enough has been set aside by the firm in investment accounts to meet the retirees' scheduled benefits. The under -- funded pension plans can be a way by which firms seek to solve a form of the last-period problem of retired workers, especially unionized workers, whose concern for the financial stability of the firm may stop when they get their gold watch. Unions often negotiate the retirement payments and fringe benefits for unionized retirees at the same time they negotiate the pay packages for the current workers. Even when retirement benefits are fixed for retirees' lives, the retirees have an interest in the continuation of the firm, but only when the pension plans are not fully funded. When they are fully funded the retirees don't have as much of a stake in the continuation of the firms. They can reason, "Who cares what the workers get paid, we've got ours!" When the retirement plans are not fully funded, the retirees must worry that excessive wage demands by current workers can decrease the ability of the firm to fund the retirement benefits in the future and thereby meet the scheduled benefit payments. Hence, under -- funded pension plans can be a way of tempering union wage demands by giving retirees a stake in wage rates than are lower than otherwise.

The very fact that an "old" owner of a business can sell to a "young" owner also enhances the incentive of the old owner to maintain the reputation of the firm. However, once the firm is sold, there is an incentive for the old owner to allow the firm's reputation to decline, a prospect that encourages a speedy transfer of a business when the deal is closed. If the new owner can't take over the business in a timely fashion, then he or she might overcome the last-period problem simply by insuring that the old owner retains stock in the business.

Of course, the new owner might prefer to have complete control of the business once it is acquired. However, the value of the share he or she controls might be greater if the old owner retains some incentive to keep the reputation and material and human resources of the business intact between the time the sale is completed and the transfer of ownership is finalized. Otherwise, the old owner may have an incentive not only to relax on the job, but also to set up a totally new business and then raid the old company of its key employees and customers.

If the old owner retains some interest in the firm, then he or she also has an incentive to work with the new owners, giving them time to develop the required reputation for honest dealing with employees and customers and to take control of one of the more elusive business assets -- the network of contacts. The practice of keeping the old owner on after the sale of the business is common among businesses such as medical offices. Doctors first form a firm that looks and operates like a partnership, after which they finalize the sale. In all of these cases, the old owners will want to work with the new owners to make the transfer as "seamless" as possible, simply because the sale price will be higher, and the greater the chance the new owner has to establish a reputation for honest dealing and to take charge of the contacts.

Scott Cook, who in 1983 developed the widely used home-finance software package called "Quicken," the major product of Cook's firm, Intuit, Inc., which was courted for a buyout in 1994 by Microsoft. Cook eventually agreed to sell Intuit to Microsoft for $1.5 billion in Microsoft stock, 40 percent above Intuit's market price at the time. Microsoft agreed to pay a premium price for a couple of reasons. First, Bill Gates, CEO of Microsoft, saw a need to have a dominant personal finance program that could be integrated into his Microsoft Office line and that would allow him to pursue his goal of transforming the way people manage their money. The value of Intuit was greater as an integrated part of Microsoft than by itself. Second, and more importantly for the purposes of this chapter, Cook agreed to become a vice president of Microsoft and to retain an interest in the future development and use of Quicken, if Microsoft bought Intuit. This way Cook could minimize the impact of the last-period problem, and the sale of Intuit would mean that Quicken might continue to develop. The proposed buyout of Intuit eventually was terminated by the Justice Department, which threatened to sue Microsoft for antitrust violation. However, the example is still a good one not only because it involves prominent business personalities and their successful firms, but also because of the moral it illuminates: Sometimes, by selling only a part of the company, an owner can increase the value of the part that is sold, enhancing the combined value of the part that is sold and the part that is retained.

The last-period problem also helps to explain why fathers (or mothers) are so anxious for one of their sons (or daughters) to go into their business as retirement age approaches. This not only extends the life of the business, but it also increases the amount of business that can be done as the retirement age is approached, given that with the elevation of the son or daughter, the last period is then put off until some time in the future.

Why do signs on business establishments sometimes read, for example, "Sampson & Sons" or "Delilah & Daughter"? The usual answer is that the parent is proud to announce that a daughter (or son) has joined the business. That is probably often the case, but we also think it has a lot to do with the parent seeking to assure customers and suppliers that the original owner, the parent, will not soon begin to take advantage of them.

Economists David Laband and Bernard Lentz have found that the rate of occupational following within families with a self -- employed proprietor is three times greater than within other families, which suggests that proprietors have good reason -- measured in continuing the value of their companies -- to bring their children into the business that other people don't have.[8] Caterpillar, the manufacturer of farm equipment and heavy machinery, depends on its dealers to maintain customer trust and goodwill. One way Caterpillar has attempted to enhance customer trust is to set up a school to help children of dealers learn about and pursue careers in Caterpillar dealerships.[9]

---

[8] David N. Laband and Bernard F. Lentz, "Entrepreneurial Success and Occupational Inheritance Among Proprietors," Canadian Journal of Economics, Vol. 23, No. 3 (August 1990), pp. 101 -- 117.

[9] William Davidow and Michael Malone, The Virtual Corporation, (New York: Harper Collins Publishers, 1992), p. 234.

Firms commonly complain that goods delivered in the last days of the supplier's operation are of inferior quality. The problem? It may be one of the incentives, or lack thereof, that people have to deliver goods of waning quality during their last days. Bankruptcy laws can be explained in part as a means of reducing these end -- period problems.[10] They extend the potential end of the firm, and can give the firm a new lease on life and set back the last-period problem indefinitely.

Also, a firm in financial trouble can be pressed into liquidation by nervous bondholders, a fact that can exacerbate the last-period problem, given that suppliers would have to worry that nervous bondholders will encourage firms to deliver shoddy merchandise, which can make customers more nervous about dealing with the financially strapped firm. By allowing firms in financial trouble to continue operating, bankruptcy laws make it more likely that the bankrupt firms will keep up the quality of the products, and provide more motivation for suppliers to keep up honest dealing.

The Keiretsu As a Solution to the Last-period Problem

Japanese firms are renown for organizing themselves into groups of firms called *keiretsus*. *Keiretsu* members buy from one another, share information, and organize joint ventures to produce goods and services in concert with one another. The largest and best-known *keiretsu* is Mitsubishi, which has 28 core member firms and hundreds of other firms that are loosely tied to the core firms. They integrate their activities in a number of ways, not the least of which is having their headquarters close together, having the CEOs of the various firms meet regularly to exchange information, and organizing social and business clubs that are open to employees of the *keiretsu* member firms. The members often own stock in one another.

In the United States, many of the activities of any *keiretsu* would likely worry the antitrust authorities because the organization would be construed as monopolistic. No doubt, some *keiretsu* activities might indeed restrain competition in some markets, causing prices of Japanese goods to be higher than they otherwise would be (especially in the domestic market where competition from other producers from around the world might be impaired by import restrictions). The *keiretsu* might also be seen as a highly efficient means by which Japanese firms are able to make use of new technologies, quickly incorporating them into products. The Japanese have demonstrated a knack for bringing new products to market quickly.

However, we mention the *keiretsu* organizational form here only because of one of its more unheralded benefits: it is a form of business organization that seeks to solve the last-period problem. The integration of the member firms' purchases and sales and strategic plans for the future is a means by which members can assure one another that their business relationship will be enduring -- or that the member employees have minimum incentive to behave opportunistically in the short -- run and have maximum

[10]Gibbons and Murphy, "Optimal Incentive Contracts in the Presence of Career Concerns."

incentive to work with their joint future income stream in mind.[11] Being ousted from the *keiretsu* can inflict substantial costs on the opportunistic firms and their employees. Even the social gatherings of *keiretsu* employees can be construed as a means by which the employees can "bond." Here, we are not so much concerned with the "warm and fuzzy" feelings people might have from integrating their lives. Instead, we mean that by integrating their lives at the social level, employees can provide each other mutual assurance that they will live up to expectations in their business dealings, that they will not act opportunistically. The employees can lose the long -- term benefits of their social and business relationships.[12]

In short, the *keiretsu* is a clever means by which opportunistic behavior is made more costly. It seeks to reduce some of the shirking and monitoring costs of doing business, when business is done at arm's length.

Indeed, one of the more unrecognized benefits of the *firm* in general is that it does, under one "roof," what is attempted under a *keiretsu*. The firm seeks to bring people together and have them associate and work together on a continuing basis for the purpose of minimizing the last-period problem. As we noted early in the book, it's quite possible for all departments within a firm and all stages of an assembly line to be operated on a market basis, with every department and every stage of the assembly line buying from one another. However, you can imagine that such an organization of economic activity would give rise to a multitude of last-period problems, especially if there were no attempt to ensure that everyone "worked together" as something approximating a *keiretsu*.

The Japanese relatively greater use of formal and informal long -- term buyer -- supplier relationships – sometimes cited as "strategic industrial sourcing" combined with so -- called "relational contracting" -- may be partially explained by the fact that the Japanese, as commonly argued, have the required business culture, one grounded in a long -- term, future -- oriented business perspective that prescribes long -- term contracts. The Japanese may, to a greater degree than Americans and Europeans, have a pervasive sense of duty that insures that the parties will abide by any contracts that have been consummated, and the Japanese may have a greater aversion than others to ongoing contentious bargaining relationships that would be required if contracts were always up for grab by the low -- cost bidders.[13] The long -- term business relationships may also be a consequence of the growing affluence in Japan, which has elevated the importance of quality over price that, in turn, has induced large Japanese firms to work with their suppliers in an effort to enhance product quality.[14] The long -- term contracting can also be explained partially by the encouragement the Japanese government gave to the

---

[11] For an interesting discussion of the *keiretsu*, see Clyde V. Prestowitz, Jr., Trading Places: How We Allowed Japan to Take the Lead (New York: Basic Books, Inc., 1988), pp. 156 -- 166.

[12] As Clyde Prestowitz notes, "Thus the *Keiretsu* system reduces risks for the Nippon Electric Company and the other Japanese companies through the accumulation of relationships that can be counted upon to cushion shock in time and trouble" (ibid., p. 164).

[13] This explanation for long -- term contracting has been argued at length by Ronald P. Dore, Taking Japan Seriously (Stanford, Calif.: Stanford University Press, 1987).

[14] Ibid., p. 188.

creation of long -- term buyer -- supplier relationships in the past (especially during World War II) and the existing laws and legal sanctions against abusive treatment of subcontractors by their customers.[15]

But it seems to us altogether reasonable that long -- term contracting must be grounded in factors other than culture and affluence. One economic explanation may start with a recognition of the extent to which firms are integrated in Japan. The fact of the matter is that in some industries Japanese production is far less integrated into identified "firms" than, say, in the United States and other countries. In the United States and Western Europe, for example, 50 to 60 percent of the automobile manufacturing costs are incurred "in -- house." In Japanese firms, on the other hand, only 25 to 30 percent of the automobile production costs are typically incurred "in -- house," or inside Japanese firms. [16] Only 20 percent of Honda's production costs are incurred inside, which means it buys 80 percent, or $6 billion, of its inputs from outside suppliers.[17] Because of the lack of integration, Japanese firms may need to develop long -- term buyer -- supplier relationships to a much greater degree than more highly integrated firms do just to overcome the potential last-period problems, if nothing else.

Put another way, Japanese firms are able to engage in what is called strategic outsourcing, and do so competitively, *because* they are willing and able to develop long -- term working relationships. If they didn't, they would have to endure the added costs associated with the ever -- present closing of those relationships. It doesn't surprise us that many buyer -- supplier relationships in Japan give the "look and feel" of integrated firms with buyers and suppliers helping each other and investing in each other (which is what happens, to more or less degree, within unified firms).

When Honda signs a contract with a supplier, it expects the working relationship to continue for 25 to 50 years, which effectively means that the last-period problem is set back considerably.[18] Moreover, the permanence of the buyer -- supplier relationship is two -- way, with commitments on the parts of both buyers and suppliers. Buyers agree to stay with the suppliers, and vice versa, through ups and downs (at least up to a point). Hence, Honda can justify incurring the costs associated with helping its suppliers increase productivity, even provide the needed technology and specialized equipment. Moreover, such expenditures, plus investments in the specific assets of the suppliers, by Honda have the added advantage of being a *bond*, the value of which is forgone if Honda does not abide by its agreement. Managers at Honda are basically saying to suppliers, "Look at what we are doing. We are serious in our commitment. If we renege, our up -- front investment will be worth very little. We will lose our projected income stream from the investment. Because of those costs, you can count us in for the long run." Such tie -- ins aid in making the contracts self -- enforcing and durable; they help to make the long run a viable perspective.

---

[15] Ibid.
[16] As reported in Toshihiro Nishiguchi and Masayoshi Ikeda, "Suppliers' Process Innovation: Understated Aspects of Japanese Industrial Sourcing," in Managing Product Development, edited by Toshihiro Nishiguchi (New York: Oxford University Press, 1996), pp. 206 -- 230.
[17] As reported in Lisa H. Harrington, "Buying Better," Industry Week, July 21, 1997, pp. 74 -- 80.
[18] Ibid.

The Role of Markets

Should production be rigidly integrated as in American firms or more loosely integrated as in Japanese business consortiums? We surely cannot answer that question with the certitude that many readers will want. Japanese firms obviously gain the benefits of keeping their suppliers in a position that is marginally more tenuous and, maybe, more competitive with other potential suppliers, but they have to deal with the marginally more severe last-period problems. Many factors, which are offsetting and subject to change with the costs associated with contracting and with principal/agency problems we have discussed, are involved. We suspect that different organizational forms will suit different situations and eras (as has obviously been the case in Japan where relational contracting has not always been prevalent[19]).

Answers will come from real -- world experimentation in the marketplace. We suspect that competition will press firms to adjust their organization forms, and the inherent incentive structures, as some variation of organizational form is relatively more successful. Many American firms have had to seriously consider and, to a degree, duplicate the added organizational flexibility of Japanese firms. Why? Their management methods have obviously worked in some industries, most notably the automobile industry. It takes 17 hours to assemble a car in Japan and 25 to 37 hours to assemble a comparable car in the United States and Europe. Japanese firms can develop a new car in 43 months, whereas it takes American and European firms over 60 months, and Japanese cars come off the production lines with 30 percent fewer defects. The worst American -- made air conditioning units have a thousand defects for every defect in the best Japanese -- made units.[20]

Firm integration and relational contracting are hardly the only means of moderating last-period problems. Joint ventures, which more often than not require up -- front investments by the firms involved, can also be seen as extensions of firm efforts to reduce last-period problems, with the potential of enhancing the quality of the goods and services produced and lowering production costs. Joint ventures might lower production costs because they give rise to economies of scale and scope through the application of technology, but they also can lower production costs by lowering the potential costs associated with opportunistic behavior and monitoring. They make the future income streams of each party a function of the continuation of the relationship.

* * * * *

The "last-period" problem is nothing more than what we have tagged it, a "problem" that businesses must consider and handle. It implies costs. At the same time, firms can make money by coming up with creative ways of making customers and suppliers believe that the "last period" is some reasonable distance into the future. Failing firms have a tough time doing that, which is one explanation why the pace of

---

[19] See Toshihiro Nishiguchi, Strategic Industrial Sourcing: The Japanese Advantage (New York: Oxford University Press, 1994), chap. 2.

[20] As reported with citations to other sources by Nishiguchi, Strategic Industrial Sourcing, pp. 5 -- 6.

failure quickens when the prospects are recognized, given that customers and suppliers can be expected to withdraw their dealings as the expected date of closing approaches.

Firms that want to continue to exist have an obvious interest in making sure there is a resale market for their firm, not just the assets that might be sold separately. The owners and workers can then capture the long -- run value of their efforts to build the firm. By highlighting the last-period problem, we are suggesting that the firm resale market can boost the long -- term value of those assets simply by alerting people to the fact that the firm can continue for some time into the future. This means that those firms -- brokers -- who make a market for the sale of firms add value in a way not commonly recognized, by giving firms the prospect of longevity.

The "hollow corporation," in which everything is "outsourced," or nothing is produced directly, is sometimes viewed as the organizational ideal, given that the firm owners can rely on competitive forces to keep the prices of what they sell as low as possible. We doubt that the "hollow corporation" will ever dominate the economic landscape of any country for a simple reason that comes out of the analysis of this "Manager's Corner": The absence of the continuing association of employees under one roof would mean that the last-period problems would arise in spades. This is because the direct association of people under one roof has an unappreciated benefit: as in the *keiretsu* in Japan, the *firm* permits the creation of abiding relationships that reduce the incentive individuals have to behave opportunistically in the short run and enhance their incentives to work with their long -- term goals in mind. "Bonding" is something that firms do.

Concluding Comments

The concept of rational behavior means that the individual has alternatives, can order those alternatives on the basis of preference, and can act consistently on that basis. The rational individual will also chose those alternatives whose expected benefits exceed their expected costs.

Traditionally economics has focused on the activities of business firms, and much of this book is devoted to exploring human behavior in a market setting. The concept of rational behavior can be applied to other activities, however, from politics and government to family life and leisure pursuits. No matter what the activity, we all tend to maximize our well -- being. Any differences in our behavior can be ascribed to differences in our preferences and in the institutional settings, or constraints, within which we operate.

Institutional settings affect people's range of alternatives and thus the choices they make. It makes sense to examine the constraints of institutional settings. In this part of the book we will investigate the specific characteristics of the market system, the subject of microeconomic theory. Later we will look at the constraints of government. In both cases the range of choices open to individuals affects the ability of the system to produce the results expected of it.

We have also indicated in this chapter how individual rationality can give rise to a nontrivial problem for managers, the last-period problem, which can make deals costly. At the same time, we have indicated how thinking in terms of rational precepts can suggest ways managers can deal with their last-period problems to lower firm costs and raise firm profitability.

Review Questions

1. What are the costs and benefits of taking this course in microeconomics? Develop a theory of how much a student can be expected to study for this course. How might the student's current employment status affect his or her studying time?
2. Some psychologists see people's behavior as determined largely by family history and external environmental conditions. How would "cost" fit into their explanations?
3. Why not base a course on an assumption of widespread "irrational" behavior?
4. Okay, so no one is totally rational. Does that undermine the use of "rational behavior" as a means of thinking about markets and management problems?
5. How could drug use and suicide be considered "rational"?
6. If your firm were consistently dealing with "irrational behavior" among the owners and workers, what would happen to correct the problem? More to the point, what might you do to correct the problem?
7. Develop an economic explanation for why professors give examinations at the end of their courses. Would you expect final examinations to more necessary in undergraduate courses or MBA courses? In which classes – undergraduate or MBA – would you expect more cheating?

CHAPTER 4

Government Controls: How Management Incentives Are Affected

Without bandying jargon or exhibiting formulae, without being superficial or condescending, the scientist should be able to communicate to the public the nature and variety of consequences that can reasonable be expected to flow from a given action or sequence of actions. In the case of the economist, he can often reveal in an informal way, if not the detailed chain of reasoning by which he reaches his conclusions, at least the broad contours of the argument.

E. J. Mishan

Earlier chapters showed how the models of competitive and monopolistic markets illuminate the economic effects of market changes, such as an increase in the price of oil. This chapter will examine the use of government controls to soften the impact of such changes. We will consider four types of government control: excise taxes, price controls, consumer protection laws, and minimum-wage laws. As we will see, government controls can inspire management reactions that negate some of the expected effects of the controls.

Who Pays the Tax?

Most people are convinced that consumers bear the burden of excise (or sales) taxes. They believe producers simply pass the tax on to consumers at higher prices. Yet every time a new (or increased) excise tax is proposed producers lobby against it. If excise taxes could be passed on to consumers, firms would have little reason to spend hundreds of thousands of dollars opposing them. In fact, excise taxes do hurt producers.

Figure 4.1 shows the margarine industry's supply and demand curves, S_1 and D. In a competitive market, the price will end toward $P2$ and the quantity sold toward Q_3. If the state imposes a $0.25 tax on each pound of margarine sold and collects the tax from producers, it effectively raises the cost of production. The producer must now pay a price not just for the right to use resources, such as equipment and raw materials, but for the right to continue production legally. The supply curve, reflecting this cost increase, shifts to S_2. The vertical difference between the two curves, P_2 and P_1, represents the extra $0.25 cost added by the tax.

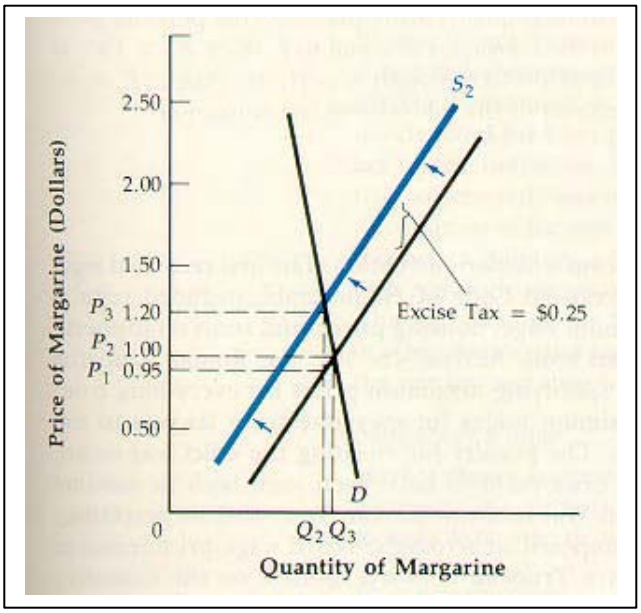


Figure 4.1 The Economic Effect of an Excise Tax

An excise tax of $0.25 will shift the supply curve for margarine to the left, from S_1 to S_2. The quantity produced will fall from Q_3 to Q_2; the price will rise from P_2 to P_3. The increase, $0.20, however, will not cover the added cost to the producer, $0.25.

Given the shift in supply, the quantity of margarine produced falls to Q_2 and the price rises to P_3. Note, however, that the price increase (P_1 to $P_{2)}$ is less than the vertical distance between the two supply curves (P_2 to P_1). That is, the price increases by less than the amount of the tax that caused the shift in supply. Clearly, the producer's net has fallen. If the tax is $0.25, but the price paid by consumers rises only $0.20 ($1.20 - $1.00), the producer loses $0.50. It now nets only $0.95 on a product that used to bring $1.00. In other words, the tax not only reduces the quantity of margarine producers can sell, but makes each sale less profitable.

Incidentally, butter producers have a clear incentive to support a tax on margarine. When the price of margarine increases, consumers will seek substitutes. The demand for butter will rise, and producers will be able to sell more butter and charge more for each pound.

The $0.25 tax in our example is divided between consumers and producers, although most of it ($0.20) is paid by consumers. Why do consumers pay most of the tax? Consumers bear most of the tax burden because consumers are relatively unresponsive to the price change. The result, as depicted in Figure 4.1, is that consumers bear most of the tax burden while producers pay only a small part (20 percent) of the tax. If consumers were more responsive to the price change, then a greater share of the tax burden would fall on producers who would then have more incentive to oppose the tax politically. Indeed, we should that the amount of money producers would be willing to spend to oppose taxes on their product (through campaign contributions or lobbying) will depend critically on the responsiveness of consumers to a price change. The more responsive consumers are, the more producers should be willing to spend to oppose the tax.

Price Controls

Price controls are by no means a modern invention. The first recorded legal code, the four-thousand-year-old Code of Hammurabi, included regulations governing the maximum wage, housing prices, and rents on property such as boats, animals, and tools. And in A.D. 301, the Roman Emperor Diocletian issued an edict specifying maximum prices for everything from poultry to gold, and maximum wages for everyone from lawyers to the cleaners of sewer systems. The penalty for violating the edict was death. More recently, wage and price controls have been used both in wartime (during the Second World War and the Korean War) and in peacetime. President Richard Nixon imposed an across-the-board wage-price freeze in 1971. Prime Minister Pierre Trudeau imposed controls on the Canadian economy in 1975. President Jimmy Carter controlled energy prices in 1977 and later proposed the decontrol of natural gas.

Wage and price controls are almost always controversial. Like attempts to control expenditures, they often create more problems than they solve. We will examine both sides of the issue, starting with the argument in favor of controls.

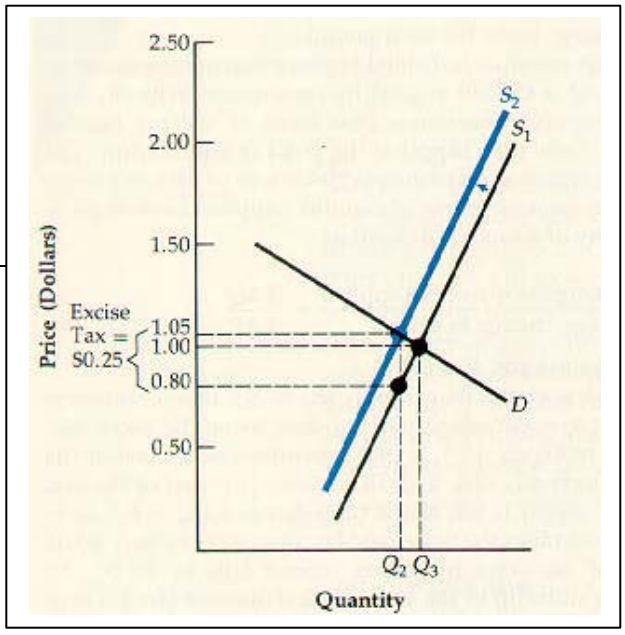


Figure 4.2 The Effect of an Excise Tax When Demand is More Elastic Than Supply

If demand is much more elastic than supply, the quantity purchased will decline significantly when supply decreases from S_1 to S_2 in response to the added cost of the excise tax. Producers will lose $0.20; consumers will pay only $0.05 more.

The Case for Price Controls

The case for price ceilings on particular products is complex. On the most basic level, many people believe that prices should be controlled to protect citizens from the harmful effects of inflation. When prices start to rise, redistributing personal income and disrupting the status quo, it seems unfair. Price controls may seem especially legitimate to people, like the elderly, who must live on fixed incomes, and have little means of compensating for the effects of price increases on goods like oil and gas.

Unearned Profits

Many proponents of price controls view the supply curve for a controlled good as essentially vertical. They believe that a price rise will not affect the quantity produced. Consumers will get nothing more in the way of goods, but producers will reap a windfall profit. Instead of an incentive to produce more, profit is seen as an economic rent—an exploitative surplus received by companies fortunate enough to be in the market at the right time.

Administered Prices

A technical argument for price controls is most often advanced by economists and public officials. Many economists maintain that a significant segment of the business and industrial community—the larger firms that control a sizable portion of industry sales—no longer responds to the forces of supply and demand. Firms in highly concentrated industries like steel, automobiles, computers, and tobacco can override market forces by manipulating their output so as to set price levels. Furthermore, they can manage the demand for their products through advertising campaigns. With market forces ineffective, control must come from the government. Price controls are the only way to avoid the production inefficiencies and inequitable distribution of income that result from concentration of industry. As John Kenneth Galbraith, a leading advocate of price controls, has put it, "Controls are made necessary because planning has replaced the market system. That is to say that the firm and the union have assumed the decisive power in setting prices and wages. This means that the decision no longer lies with the market and thus with the public."[1]

Monopoly Power

Later in the course, we will see how a monopolist can be expected to restrict output in order to push up its price in order to earn greater profits. The case for price controls under monopoly conditions is, for many advocates of controls, a matter of "fairness." The controls give back to consumers what they "deserve" in terms of lower prices. However, as we will see, under monopoly conditions, if the producer is forced to charge a (somewhat) lower price, the producer will rationally choose to increase the output level. Hence, price controls benefit consumers in two ways, first through lower prices and then through greater output.

The Case Against Price Controls

Just as the case for price controls is tied closely to the existence of monopoly power, the case against controls rests heavily on the competitive market model. Economists who

[1] John Kenneth Galbraith, *Economics and the Public Purpose* (Boston: Houghton Mifflin, 1973), p. 315.

oppose controls feel that competition is sufficient to govern business behavior, including pricing decisions. Opponents of controls also stress the individual's right to act without government interference—a right they see as crucial to a society's ability to adjust to social and environmental change.

When we say that the prices of certain products should be controlled by government, what do we mean by "government"? Can government as we know it consistently reflect the public interest? Is government immune to human failings? Opponents of price controls emphasize that the pricing decisions made by any government agency will reflect the will of its staff. Personal preference will loom large in their decisions on what constitutes a just price and a just allocation of goods and services. Political considerations may also play a role. Firms with a talent for political maneuvering will have an advantage under a price control system. In other words, competitive behavior is not necessarily reduced by price controls, though its form of expression may be changed.

If price controls are complemented by a system of government allocation of supplies, then strikes, demonstrations, and violence may also influence government decisions. During the energy crisis of 1973—1974, and again in 1978, the federal government regulated the allocation of crude oil between gasoline and diesel fuel producers. When truckers received less fuel than they claimed they needed, independent drivers stuck, threatening to paralyze the nation's commerce unless they got more fuel at lower prices. To ensure cooperation among drivers, the strikers blocked roads, vandalized the equipment of nonstrikers, and shot at drivers who ventured out on the road. One trucker was killed, and others were seriously injured. At least for a short time, such tactics were productive. The government agreed to earmark more crude oil for diesel fuel production and to lower the federal excise tax on diesel fuel. (Courts later declared those decisions illegal.)

Shortages and the Effective Price of a Product

In a competitive market, any restriction on the upward movement of prices will lead to shortages. Consider Figure 4.3, which shows supply and demand curves for gasoline. Initially, the supply and demand curves are S_1 and D, and the equilibrium price is P_1. Now suppose that the supply of gasoline shifts to S_2, and government officials, believing that the new equilibrium price is unjust, freeze the price at P_1. What will happen to the market for gasoline?

At price P_1, which is now below equilibrium, the number of gallons demanded by consumers is Q_2, but the number of gallons supplied is much lower, Q_1. A shortage of Q_2 -- Q_1 gallons has developed. As a result, some consumers will not get all the gasoline they want. Some may be unable to get any.

Because of the shortage, consumers will have to wait in line to get whatever gasoline they can. To avoid a long line, they may try to get to the service station early—but others may do the same. To assure themselves a prime position, consumers may have to sit at the pumps before the station opens. In winter, waiting in line may mean wasting gas to keep warm. The moral of the story: although the pump price of gasoline may be

held constant at P_1, the effective price -- the sum of the pump price and the values of time lost waiting in line -- will rise.

Shortages can raise the effective price of a product in other ways. With a long line of customers waiting to buy, a service station owner can afford to lower the quality of his service. He can neglect to clean windshields or check oil levels, and in general treat customers more abruptly than usual. As a result, the effective price of gasoline rises still higher. Again, during he energy crises of 1973-1974 and 1978, some service station owners started closing on weekends and at night. A few required customers to sign long-term contracts and pay in advance for their gasoline. The added interest cost of advance payment raised the price of gasoline even higher.

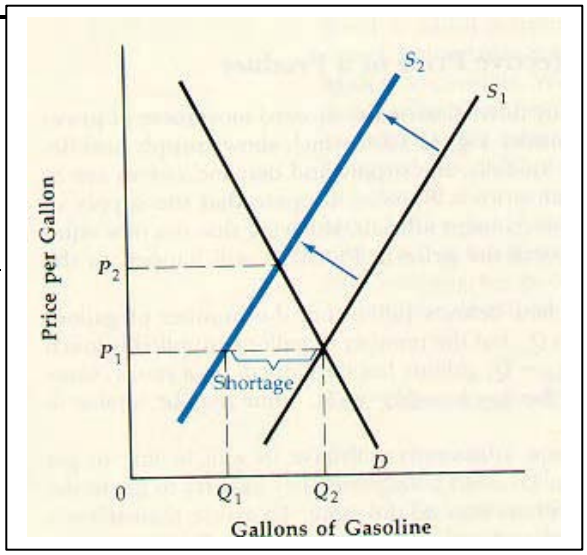


Figure 4.3 The Effect of Price Controls on Supply

If the supply of gasoline is reduced from S_1 to S_2, but the price is controlled at P_1, a shortage equal to the difference between Q_1 and Q_2 will emerge.

Black Markets and the Need for Rationing

Besides such legal maneuvers to evade price controls, some businesses may engage in fraud or black marketeering. During the winter of 1973—1974, a good many gasoline station owners filled their premium tanks with regular gasoline and sold it at premium prices. At the same time, a greater-than-expected shortage of heating oil developed. Truckers, unable to get all the diesel fuel they wanted at the controlled price, had found they could use home heating oil in their trucks. They paid home heating oil dealers a black market price for fuel oil, thus reducing the supply available to homeowners. As always, government controls bring enforcement problems.

To assure fair and equitable distribution of goods in short supply, some means of rationing is needed. If no formal system is adopted, supplies will be distributed on a first-come, first-served basis—in effect, rationing by congestion. A more efficient method is to issue coupons that entitle people to buy specific quantities of the rationed good at the prevailing price. By limiting the number of coupons, government reduces the demand for the product to match the available supply, thereby eliminating the shortage and relieving the congestion in the marketplace. In Figure 4.4, for example, demand is reduced from D_1 to D_2.

The coupon system may appear to be fair and simple, but how are the coupons to be distributed? Clearly the government will not want to auction off the coupons, for that would amount to letting consumers bid up the price. Should coupons be distributed equally among all consumers? Not everyone lives the same distance from work or school. Some, like salespeople, must travel much more than others. Should a commuter receive more gas than a retired person? If so, how much more? Should the distribution of coupons be based on the distance traveled? (And if such a system is adopted, will people lie about their needs?) These are formidable questions that must be answered if a coupon system is to be truly equitable. By comparison, the pricing system inherently allows people to reflect the intensity of their needs in their purchases.

Once the coupons are distributed, should the recipients be allowed to sell them to others? That is, should legal markets for coupons be permitted to spring up? If the deals made in such a market are voluntary, both parties to the exchange will benefit. The person who buys coupons values gasoline more than her money. The person who sells his coupons may have to cut back on driving, but he will have more money to buy other things. The seller must value those other things more than lost trips, or he would not agree to make the exchange. The positive (and often high) market value of coupons shows that price controls have not really eliminated the shortage.

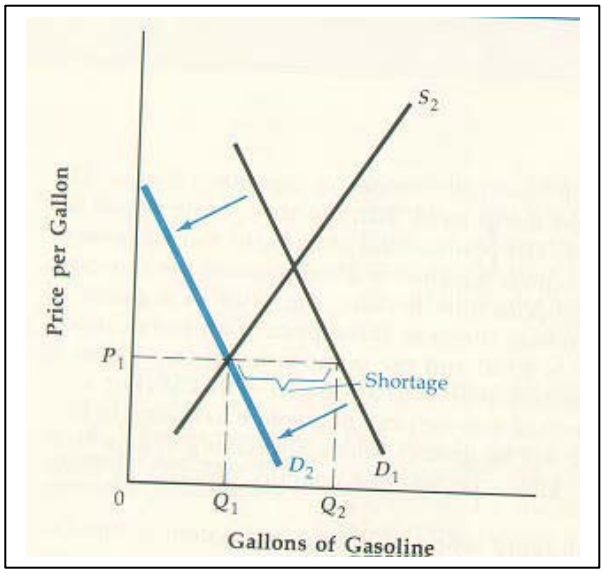


Figure 4.4 The Effect on Rationing on Demand

Price controls can create a shortage. For instance, at the controlled price P_1, a shortage of Q_2 -- Q_1 gallons will develop. By issuing a limited number of coupons that must be used to purchase a product, government can reduce demand and eliminate the shortage. Here rationing reduces demand from D_1 to D_2, where demand intersects the supply curve at the controlled price.

Furthermore, if the coupons have a value, the price of a gallon of gasoline has not really been held constant. If the price of an extra coupon for one gallon of gasoline is $0.50 and the pump price of that gallon is $1.25, the total price to the consumer is $1.75 ($0.50 + $1.25). The existence of a coupon market means that the price of gasoline has risen. In fact, the price to the consumer will be greater under a rationing system than under a pricing system. This is because the quantity supplied by refineries will be reduced.

Perhaps the most damaging aspect of a rationing system is that the benefits of such a price increase are not received by producers—oil companies, refineries, and service stations—but by those fortunate enough to get coupons. Thus the price increase does not provide producers with an incentive to supply more gasoline. (If the increase went to producers, their higher profits would encourage them to search for new sources of oil and step up their production plans.)

Consumer Protection

Less than one hundred years ago the general rule of the marketplace was caveat emptor—"let the buyer beware." The individual consumer was held responsible for the safety, quality, and effectiveness of his purchases. The seller could assume liability for the safety and effectiveness of goods and services, but only through a contract endorsed by both parties. The same rule applied to contracts: the buyer was responsible for what he signed. Although consumers could sue sellers for breach of contract or for fraud, no government agency would initiate the suit. Nor did government protect citizens in other ways from the products they bought.

During this century, however, product liability has gradually shifted from the consumer to the producer and the seller. Both court decisions and changes in the law have contributed to this shift. Many now see consumer protection as a government function.

The Case for Consumer Protection

The argument for relieving consumers of product liability resembles the argument for regulation of utilities in many respects. Both cases hinge on the costs of gaining information and the problems created by external benefits and costs and monopoly power.

External Benefits

When two cars collide, both cars will sustain less damage and both drivers less injury if just one of the cars is equipped with protective bumpers. Thus people who do not buy protective bumpers can benefit from others' investments. If many car buyers ignore the benefits others may receive from their purchases, the quantity of shock-absorbing bumpers sold will be less than the socially desirable or economically efficient amount.

This analysis of external benefits can be extended to include the concept of consumer protection. Suppose the supply curve in Figure 4.5 is the industry's willingness to offer protective bumpers. The demand curve D_1 represents consumer demand based on the private benefits to consumers, while D_2 represents private plus public (external) benefits. Under competitive conditions, the quantity produced and sold in the marketplace will be Q_1—even though up to Q_2, the total benefits of bumpers exceed their cost. The private benefits of the bumpers are small enough that many people cannot justify purchasing them.

Graphically, the vertical distance between the two demand curves, *ab*, represents the external benefits per bumper sold that are not being captured by the market. Government can close this gap by setting product standards. By requiring new cars to have shock-absorbing bumpers, government effectively increases demand from D_1 to D_2. It forces people to expand their purchases from Q_1 to Q_2, thus capturing the external benefits shown by the shaded area *abc*.

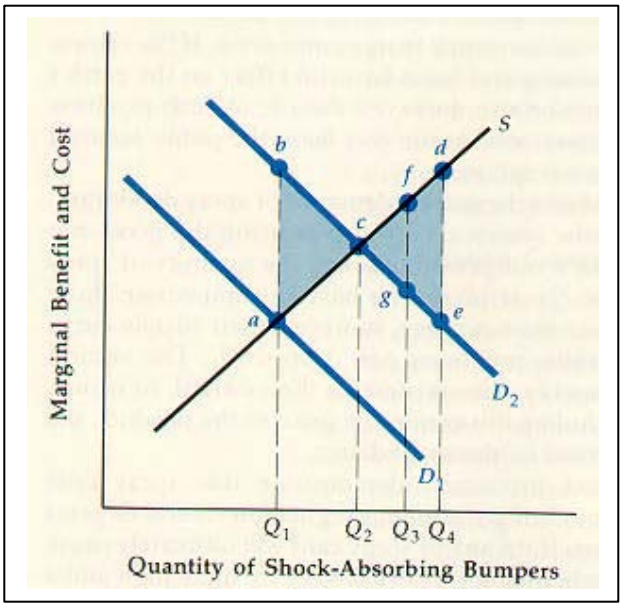


Figure 4.5 The External Benefits of Consumer Protection

Private demand for shock-absorbing bumpers is shown by the demand curve D_1: total demand (private plus public, or external, benefits), by D_2. The vertical distance between the two curves represents the social benefits from each bumper. In a free market, Q_1 bumpers will be sold. If *all* benefits are considered, however, the efficient output level will be Q2. By requiring people to purchase Q_2 bumpers, government can capture the external benefits shown by the shaded area *abc*. If the government requires consumers to buy more than Q_2 bumpers, however, excess costs will be incurred. If Q_4 bumpers are purchased, their excess social cost, shown by the shaded area *cde*, will offset their social benefits (*abc*). The *net* social gain will be zero.

This approach can be extended to a wide range of goods and services that offer significant external benefits, from safety caps for drugs to protective devices for explosives. This argument does not justify unlimited government intervention, however. We cannot conclude, for example, that all automobiles must have shock-absorbing bumpers. Such a requirement might result in the purchase of far more than Q_2 bumpers. Beyond Q_2, the marginal cost of safety bumpers is greater than their marginal benefit. An excess burden, or net social cost, is incurred when the public must purchase more than Q_2.

If the public is required to purchase Q_4 bumpers, for instance, the excess burden will be equal to the shaded area *cde*. The social cost of extending purchases to Q_4 just equals the social benefits of extending purchases to Q_2 (shown by the area *abc*). Consequently, there is no real net social benefit in moving to Q_4. If the required number of bumpers is greater than Q_2 but less than Q_4, some net social benefit will be realized. At Q_3, the excess social cost *cfg* is smaller than the social benefit *abc*. Some net benefit will be realized.

Up to a point, then, consumer protection can be socially beneficial. Society, however, can end up purchasing too much of a good thing. It is possible to make the world so safe that few resources are available for any other purpose.

Nevertheless, governments tend to require safety devices for all products in a category. Determining the optimum quantity is so difficult and costly that a blanket rule is preferable. Yet as opponents of consumer protection point out, the blanket rule itself may be extremely costly if it requires more than the socially beneficial quantity to be produced. Ultimately, the question comes down to the actual costs and benefits of particular product standards.

External Costs

The argument for consumer protection based on external costs is closely related to the argument for pollution control (a point to be taken up later in the book). If the consumers who use a product do not bear all the costs associated with its use, they will tend to consume more of the good than is socially desirable. In the process they will impose a cost on others. For example, a person who buys a spray deodorant incurs a private cost equal to its money price. If the release of the chemicals used in aerosol sprays has a harmful effect on the earth's ozone layer, as many scientists believe, however, the use of such products imposes an external cost on nonusers. At the very least, the public incurs a risk cost from the use of aerosol sprays.

Curve D in Figure 4.6 shows the market demand for spray deodorant. The supply curve S_1 shows the marginal cost of producing the good, not counting the ozone effect. In a competitive market, the quantity of spray deodorant purchased will be Q_2. If producers have to compensate those who bear the external costs of their product, however, their supply curve will shift to S_2, and the quantity purchased will drop to Q_1. The vertical distance between the two supply curves represents the external, or ozone, cost of each can sold. By including this cost in the price of the product, the government reduces social costs by the shaded area.

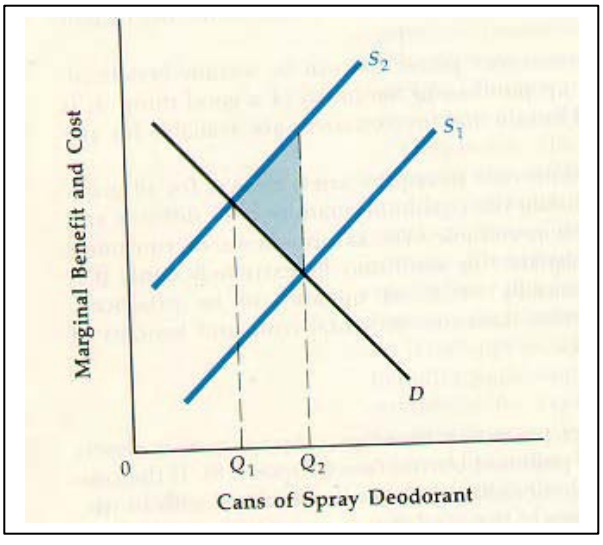


Figure 4.6 The External Costs of Consumer Protection

Curve S_1 represents the supply curve for spray deodorant, not including external costs. Curve S_2 represents the total cost, including harm to the earth's ozone layer. Thus the vertical distance between S_1 and S_2 shows the external cost of producing each can of spray deodorant. In a free market, Q_2 cans will be produced—more than the efficient level, Q_1. Government can eliminate over-production by internalizing the external costs of production, shown by the shaded area.

The argument does not necessarily demonstrate that spray cans should be banned. The amount of government regulation should depend on the degree of external cost. If

the use of spray cans will ultimately cause the destruction of life on earth, then the external costs are quite high and a complete ban is in order. If costs are lower, a less stringent policy might be appropriate.

Monopoly Power

Consumer advocates suspect that some firms use their market power to restrict the variety of products available to consumers and to reduce their quality, safety and effectiveness. The monopolist, in other words, can choose not only what price and quantity of a given product to offer, but what features it will have. Left to itself, the monopolistic firm will maximize profits by finding that one combination of price, quantity, and product features that minimizes costs and maximizes revenues.

Consumer advocates argue that most consumers want safer, more effective products than they can now obtain and are willing to pay competitive prices for them. They see consumer protection laws as a means of forcing monopolistic producers to provide what the public wants.

Information Costs

The complexities of modern technology can be overwhelming. Proponents of consumer protection argue that consumers cannot hope to comprehend the ins and outs of the dozens of products they must consider, from color televisions to prescription drugs. For instance, the production of cereals and meats is so far removed from common experience that consumers have little idea what chemicals may be added during processing. Without adequate information and the technical ability to comprehend it, consumers cannot make rational choices based on true costs and benefits. Therefore product safety experts must protect them.

This line of reasoning resembles the argument for a standard requiring shock-absorbing bumpers. Like the bumpers, consumer information benefits far more people than those who pay for it. That is, there are external benefits associated with its provision. The market demand for information, like the market demand for protective bumpers, will not fully reflect its social benefits. Because of external benefits, the quantity of information produced and purchased will fall short of the efficient level. By intervening in the market to supplement the information flow, government can increase social welfare.

The Case Against Consumer Protection

Some of the arguments against consumer protection have already been mentioned. In this section we will reemphasize them and highlight some additional points. As these arguments and counter arguments suggest, consumer protection is a complex issue, and it is difficult to find an efficient solution to the problem.

Competition as a Form of Consumer Protection

No one can reasonably expect to be protected against all the whims and exploitative efforts of businesses. The cost of complete protection would be prohibitive, and the benefits often too small to justify the cost. Thus we should not expect the market system to protect consumers completely against unsafe products and services. The relevant question is whether the market or government is more efficient in accomplishing the task of consumer protection.

In answering that question it is important to remember that few consumers are as powerless as consumer protection advocates maintain. Although one person can do little to coerce producers into providing safer, more effective goods and services, collectively consumers have considerable power of persuasion. They can offer to pay more for a safer product—and there is some price that profit-maximizing entrepreneurs will accept for such a product—or they can turn to different producers to obtain what they want. If producers do not offer what consumers want, and if they repeatedly produce shoddy merchandise, more and more consumers will move to other producers or purchase substitute goods. For example, if the Coca-Cola Company persisted in selling drinks that had lost their carbonation, consumers could move to Pepsi, 7-Up, or other substitute drinks. The fear of losing customers helps keep producers in line, pressuring them to offer the goods customers want.

Differences in Risk Taking

Some people are more willing than others to assume the risk that goods and services may be defective, ineffective, or unsafe. They differ in the personal value they place on avoiding risk. Thus some will participate in dangerous sports like hang gliding, while others would not dare. Some people will take a chance on buying a used car or toaster, while others would always insist on (and pay more for) new merchandise. If surveys are correct, most drivers are willing to accept the risk of driving without seat belts—although a few would not go around the block without them. Everything else held equal, people with a strong aversion to risk will demand safer products than those who prefer to take their chances.

Such differences in the willingness to assume risk may reflect differences in economic circumstances. Some believe that the demand for safer products is positively related to income. The rich are far more likely to buckle their seat belts than the poor. Even the choice of restaurants by the rich and poor may reflect different attitudes toward risk. People with low incomes patronize greasy-spoon restaurants, accepting the risk of food poisoning. They may reason that they are better off by eating cheaply than by spending more to protect their health.

If all consumers were willing to accept the same degree of risk, it would be relatively easy to protect them through product standards. Government regulators would simply determine the level of risk acceptable to all, and set their standards accordingly. Of course, consumer choice would be restricted. Some ineffective or less safe products would no longer be offered for sale. In the real world, as we have observed, consumers differ in their risk aversion. Uniform standards would force those who are comparatively

efficient in coping with risk, or who have no real aversion to risk, to buy safer products. Assuming that safety is not a free good, the cost to consumers would increase—and in economic terms, that amounts to a misallocation of resources. People who do not have children, for example, must still pay for childproof caps on drug bottles.

If full liability for product safety and effectiveness were shifted to the producer, the same type of problem could develop. Again, consumers would be unable to choose their preferred level of risk. When producers assume the risk, they might decide to discontinue certain product lines to avoid lawsuits and damage claims, or they might buy insurance to cover their newly acquired risk cost, raising the price to the consumer. In effect, consumers would be forced to buy insurance against unsafe or defective products. They would no longer have the option of insuring themselves, perhaps at a lower price.

The Needs of the Poor

Many people support consumer protection because of their concern for the poor, who may be unable to afford the information necessary to make an informed choice. The poor may also be the least capable of understanding technical product information, and the least able to endure the losses associated with defective goods and services. Opponents of consumer protection point out that the poor often prefer to buy low-quality goods and services because they are less expensive. They pay less so they can have more of other goods and services. If less safe (but cheaper) products are removed from the shelves, then, the burden of consumer protection falls disproportionately on the poor.

MANAGER'S CORNER: **The Importance of Manager Incentives in the Minimum-Wage Debate**

Political support in Congress for another hike in the federal minimum wage is growing. Following the lead of President Clinton, who called for an increase in the minimum wage in his 1999 State of the Union message, Senator Edward Kennedy (D-Mass.) and Representative David Bonior (D-Mich.) have proposed that the minimum hourly wage be raised by $1, or from $5.15 currently to $6.15 in two steps over the next year and a half.[2]

Indeed, even Republican members of Congress appear ready to press for their own increase in the minimum wage this year. Representative Jack Quinn (R-NY) has argued, "I believe it is a forgone conclusion that some type of minimum wage increase bill will be approved in this session of Congress. Rather than fight the thing and have Republicans being dragged kicking and screaming to a vote on the minimum wage, I say to my party, 'Why not take the lead?'"[3] Other political interest groups will draw on the support of many members of Congress in their effort to defeat any proposed increase.

---

[2] The Kennedy and Bonior companion bills would, if passed, raise the minimum wage from $5.15 to $5.65 on September 1, 1999 and to $6.15 an hour on September 1, 2000 [House, U.S. Congress, 106th Cong., 1st Sess., "Fair Minimum Wage Act of 1999," H.R.325 (January 19, 1999); Senate, U.S. Congress, 106th Cong., 1st Sess., "Fair Minimum Wage Act of 1999," S. 192 (January 19, 1999)].

[3] As quoted by Janet Hook, "GOP Relaxes Opposition to Minimum Wage Increase Politics: Republican Leaders Hope to Head Off Campaign. Hike May Be Tied to Tax Cuts," Los Angeles Times, April 12,

Both sides to the heated debate that is also unavoidable will once again restate old and tired arguments, and they both will be off course in their arguments. In considering a new round of minimum wage increases, both minimum wage proponents and opponents need to reconsider how a minimum-wage hikes will affect labor market incentives and manager reactions to what Congress legislates. By the same token, managers in markets affected by any new minimum-wage increase need to be mindful of the competitive forces afoot that will cause them to react to an increase in ways that they might not always like.

The History of the Minimum Wage in Current and Constant Dollars

In emerging debate, much will likely be made of how the current federal minimum wage of $5.15 an hour has no more purchasing power than the minimum wage of the early 1950s, a fact that can be seen in Figure 4.7. The chart shows that the minimum wage in current dollars has risen in a series of nineteen steps from 25 cents an hour when the first federal minimum wage took effect in October 1938 to $5.15 currently. However, in constant, (February) 1999 dollars the minimum wage rose irregularly from $2.92 an hour in October 1938 to $7.70 an hour in 1968, only to fall irregularly from the 1968 peak to its current level of $5.15, which is a third less than the 1968 peak. As can also be seen, the real value of the 1999 minimum wage was slightly below the real minimum wage when it was raised at the start of 1950 (at which time it was $5.25 in 1999 dollars). In recent years, the real minimum wage has fallen only slightly in real terms from $5.25 in October 1997, at which time the minimum wage was last raised, to $5.15.[4]

The Two Sides to an Old Debate

When the next minimum-wage bill reaches the floor of Congress, it is all but certain that many opponents and proponents in and out of Congress will once again lock political horns over the proposal, no matter what the proposed increase is. While the political partisans can be expected to repeat past claims in earnest, they all will once again be off base on the likely employment consequences of the minimum-wage increase.

1999, p. A1. Quinn's bill would delay the full $1 increase until September 1, 2001, but it would go one step further and raise the minimum wage annually by the consumer price index after September 1, 2002 [House, U.S. Congress, 106th Cong., 1st Sess. "Long Term Minimum Wage Adjustment Act of 1999," H.R. 964 (March 3, 1999).

[4] Over the past six decades, the percent of nonsupervisory workers covered by the federal minimum wage has risen from 57 percent in 1950 to 87 percent in 1988 (the latest year of available data). This rise in the coverage of the minimum wage should have led to any increase in the minimum wage to have a progressively greater negative employment effect over the years, which is what economist Marvin Kosters has found [Marvin H. Kosters, Jobs and the Minimum Wage: The Effect of Changes in the Level and Pattern (Washington, D.C.: American Enterprise Institute, 1989), p. A-13].

House Majority Leader Dick Armey, a long-time opponent of the minimum wage, has already declared that the proposed $1 increase in the minimum wage is the "wrong thing" to do, mainly because the increase would significantly reduce employment of the country's low skilled workers.[5] No doubt, Armey is thinking in terms of a supply-and-demand model that he once taught in his economics classes at North Texas State University. Consider Figure 4.8. If the market is competitive and free of government intervention, the wage rate will settle at W_1. Suppose, however, that politicians consider that market wage too low to provide a decent living. They pass a law requiring employers to pay no less than W_2. The effect of the law will be to reduce employment. Employers will not be able to afford to employ as many people, and the quantity of labor demanded will fall from Q_2 to Q_1. Those who manage to keep their jobs at the minimum wage will be better off; their take-home pay will increase. Other workers may no longer have a job. The will either become permanently unemployed or settle for work in a different, less desirable labor market. If the minimum wage displaces them from their preferred employment to their next-best alternative, their full wage rate—that is, their money wage plus the nonmonetary benefits of their job—will have been reduced. If they become permanently unemployed, their money wage will have been reduced from a level judged politically unacceptable to zero.

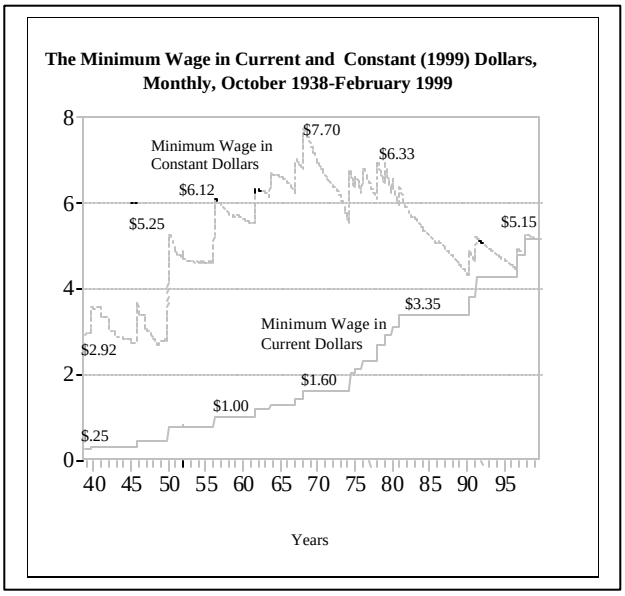


Figure 4.7 The History of the Minimum Wage in Current and Real Dollar Terms

The minimum wage rose in current dollars from $.25 an hour in 1938 to $5.15 until late 1999. However, in real (1999) dollars, the minimum wage rose from $2.92 in 1938 to $7.70 in 1968, only to fall back to $5.15 an hour in 1999.

To make matters worse, the introduction of a minimum wage increases the number of laborers willing to work (see Figure 4.8). Thus the workers who would have had a job at W_1, and who have fewer employment opportunities at W_2, must now compete

[5] "U.S. Republicans Concede GOP Support for Minimum Wage Boost," Dow Jones News Service, 1999 (as found on the Dow Jones Interactive Publication Library, April 28, 1999).

with a larger number of workers. Indeed, many of these new arrivals to the market will take jobs once held by menial workers at the market-clearing wage, W_1.

On the other side of the argument, Bob Herbert, a columnist for the New York Times and a minimum-wage supporter, approvingly quotes a study from the Economic Policy Institute, a Washington, D.C.-based think tank, that found the last approved minimum-wage hike raised the incomes of 10 million Americans.[6] Herbert writes, "The benefits of the increase disproportionately help those working households at the bottom of the income scale. Although households in the bottom 20 percent (whose average income was $15,728 in 1996) received only 5 percent of total national income, 35 percent of the benefits from the minimum wage increase went to these workers. In this regard, the increase had the intended effect of raising the earnings and incomes of low-wage workers and their households.''[7] Moreover, in the growing debate proponents like Herbert will continue to cite statistical studies that show that a minimum wage hike will have no (or minimal) impact on the count of low-wage jobs, which is what the Economic Policy Institute study found.[8]

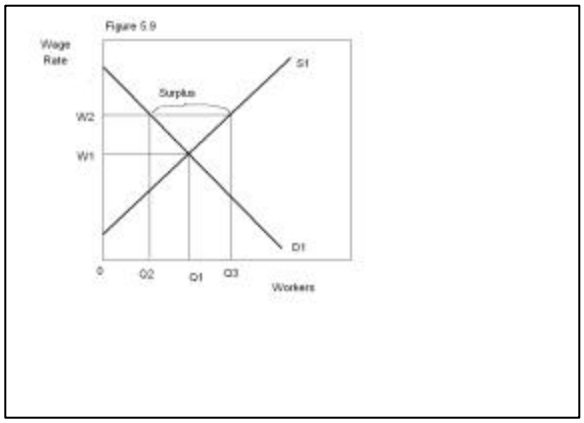


Figure 4.8 The Standard View of the Minimum Wage

When Congress raises the minimum wage from W_1 to W_2, the number of workers hired goes from Q_1 to Q_2, while the number of workers who are willing to work goes from Q_1 to Q_3. The result is a "surplus" of workers equal to $Q_3 - Q_1$. Some workers gain at the expense of others.

Herbert is convinced that such findings should give minimum-wage critics reason to eat their words. Herbert reminds his readers of Cato Institute's chairman William Niskanen (and former acting chairman of President Reagan's Council of Economic Advisors and opponent of minimum-wage increases) comments made in the middle of the previous debate over increasing the minimum wage, ''It is hard to explain the continued support for increasing the minimum wage by those interested in helping the working poor.''[9] Herbert and other minimum-wage supporters will point once again to

[6] Jared Bernstein and John Schmitt, "Making Work Pay" (Washington, D.C.: Economic Policy Institute, 1998, mimeographed).

[7] Bob Herbert, "In America; The Sky Didn't Fall," New York Times, June 4, 1998, p. A27.

[8] Bernstein and Schmitt, "Making Work Pay."

[9] Ibid.

the empirical work of Princeton University economists David Card and Alan Krueger who concluded in 1994 that the minimum-wage increases in the federal minimum wage in the early 1990s had no measurable negative effect on employment in New Jersey fast-food restaurants (and may have actually increased employment slightly).[10] They also insisted in 1998 insisted that more recent employment data from the Bureau of Labor Statistics corroborate their earlier findings.[11]

Nevertheless, opponents will continue to argue, as they have in the past, that if Congress raises the cost of low-skill labor, less than a fifth of the wage gains will go to households with incomes below the poverty level and more than half of the wage gains will go to households with more than twice the poverty income threshold.[12] They will also stress that several hundred thousand jobs are bound to be lost. Some employers will not be able to afford as many workers, and other employers can be expected to automate low-skill jobs out of existence. The opponents will back up their claims with their own statistical studies that will show that some low-skilled workers will be made better off (those who keep their jobs) but only because other low-skilled workers will be made worse off (those who are unemployed).[13] For example, the Employment Policies Institute, another Washington, D.C. based think tank, commissioned a study of the labor market impact of a $1.35 increase in the minimum-wage in the State of Washington and found that by 2000, the increase can be expected to destroy 7,431 jobs in the state, causing the affected workers to lose $64 million in annual income.[14]

Both sides to the debate will once again be wrong in their assessments of the minimum-wage increase because they have both failed to recognize that employers are a lot smarter and are pressed far more by the forces of their labor markets than the political combatants seem to think. Neither side seems to realize that Washington simply doesn't have the requisite power over markets to significantly improve worker welfare by wage decrees, no matter how well intended the legislation may be. This is why so many

---

[10] David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," American Economic Review, vol. 84 (1994), pp. 772-793; or David Card and Alan B. Krueger, Myth and Measurement: The New Economics of the Minimum Wage (Princeton, N.J: Princeton University Press, 1995).

[11] David Card and Alan B. Krueger, "Unemployment Chimera," Washington Post, March 6, 1998, p. A25.

[12] As reported by Kenneth A. Couch, "Distribution and Employment Impacts of Raising the Minimum Wage," FRBSF Economic Letter (San Francisco: Economic Research, Federal Reserve Bank, February 19, 1999, no. 99-06), p. 1. Couch cites Richard V. Burkhauser, Kenneth A. Couch, and Andrew J. Glenn, "Public Policies for the Working Poor: The Earned Income Tax Credit Versus Minimum Wage Legislation," Research in Labor Economics, edited by Sol Polacheck, pp. 65-110.

[13] Several recent statistical studies on the negative employment and income impacts of state and federal minimum wage hikes can be found on the Employment Policies Institute web site (http://www.epionline.org/research_frame.htm).

[14] David A. Macpherson, "The Effects of the 1999-2000 Washington Minimum-Wage Increase" (Washington, D.C.: Employment Policies Institute, May 1998, as found at http://www.epionline.org/research_frame.htm)

empirical studies show minimum wage increases have had a relatively small impact on employment. Indeed, most studies undertaken over the past three or four decades have found that a 10 percent increase in the minimum wage will lower the employment of teenagers (the group of workers most likely to be adversely affected by the minimum wage) by a surprisingly small percentage, anywhere from .5 to 3 percent,[15] and tight labor markets, which exist currently in the United States, imply relatively smaller reductions in the count of lost jobs with any given percentage increase in the minimum wage.[16] When labor economists were asked to give their personal estimate of the employment effect of a 10 percent increase in the minimum wage, researchers found that the surveyed economists estimate that teenage employment would fall by 2.1 percent.[17]

Why Minimum-Wage Hikes Don't Seem to Affect Employment Very Much

Why have the percentage estimates of job losses been so low? The simple answer is the labor markets for low-skilled workers are highly competitive, which explains the low wages paid workers with limited skills in the first place. Many employers of low-skilled workers would love to be able to pay their workers more, but they have to face a market reality: if they paid more, then their competitors would have a cost advantage in pricing their products.

When Congress forces employers to pay more in money wages, it also forces them to pay less in other forms, most notably in fringe benefits. If there are few fringes to take away, the employers can always increase work demands.

Why would employers curb benefits and increase work demands? There are three reasons:

---

[15] For reviews of the minimum-wage literature, see Charles Brown, Curtis Gilroy, and Andrew Kohen, "The Effect of the Minimum Wage on Employment and Unemployment," Journal of Economic Literature, vol. 20 (1982), pp. 487-528; and Charles Brown, "Minimum Wage Laws: Are They Overrated?" Journal of Economic Perspectives, vol. 2 (1988), pp. 133-147. In more recent studies in the 1990s, the reported employment effects among teenagers continue to be relatively small [Richard V. Burkhauser and David Whittenberg, "A Reassessment of the New Economics of the Minimum Wage Literature Using Monthly Data from the SIPP and CPS" (Syracuse, N.Y.: Center for Policy Research, Syracuse University, 1998).

[16] These estimates of the responsiveness of labor markets to minimum wage hikes are independent of the tightness of labor markets. If the country's labor markets remain relatively tight over the next year or so, the number of low-skill workers covered by the minimum wage can be expected to fall as market-determined wage rates for low-skill workers rise past the proposed new levels for the minimum wage. (Currently, only about 4 million Americans work at the federal minimum wage.) Hence, while the percentage reduction in the number of minimum wage jobs may remain more or less in line with past studies, it stands to reason that the actual number of minimum wage jobs will fall as the count of covered workers shrinks.

[17] Victor R. Fuchs, Alan B. Krueger, and James M. Poterba, "Economists' Views about Parameters, Values, and Policies: Survey Results in Labor and Public Economics," *Journal of Economic Literature*, vol. 36 (September 1998), pp. 1387-1425.

First, they can do it, given that the minimum-wage hike will attract a greater number of workers (and workers who are more productive) and cause some employers to conclude that they cannot hire as many workers -- *unless adjustments are made*. Hence, given the tightness of the labor market, the forced wage hike necessarily strengthens the bargaining position of employers, given that the employers can tell prospective workers, "If you don't like it, I can hire someone else. Your replacements are lined up at my personnel office door."[18] Employers will make the adjustments for an offensive reason, to improve their profits (or curb losses).

Second, and perhaps more importantly, employers of covered workers must (to decrease costs) cut fringes and/or increase work demands or face the threat of losing their market positions as their competitors cut fringes and increase work demands. Employers will, in other words, make adjustments for defensive reasons, to prevent their market rivals from taking a portion of their markets and causing their profits to fall (or losses to mount).

Third, if employers don't cut fringes and/or increase work demands, the value of the company's stock will suffer on the market, leaving open profitable opportunities for investors to buy the firm, change the firm's benefit/work demand policies, improve the firms profitability, and then sell the firm at a higher market value. Employers -- either the original or new owners -- will make the adjustments for financial reasons, to maximize share values.[19]

The net effect of the adjustments in fringes and work demands is that the cost impact of the minimum-wage hike will be largely neutralized. For example, when the minimum wage is raised by $1, the cost of labor may, on balance, rise by only 5 cents. Such an adjustment explains why the Card and Krueger studies and more than a hundred other statistical studies on the minimum wage have found that minimum-wage hikes have caused a small (if not negligible) percentage drop in jobs even among that group of workers – teenagers working at fast food restaurants – whose jobs are most likely to be cut.[20]

---

[18] Tight labor markets, like the ones in the United States in 1999, can cause wages and fringe benefits to rise, even for low-skill workers, and can cause the number of workers affected by any minimum wage hike to fall. However, the point that minimum wage hikes increase the relative bargaining power of employers still holds for those workers remaining at the minimum wage. Moreover, if employers have responded to their tight labor markets by increasing their workers' fringe benefits, then there will be more benefits for employers to take away when faced with a hike in the mandated money wage rate.

[19] Indeed, it may be interesting to note that, at least conceptually, minimum-wage workers might contemplate the prospects of buying their firms, if their firms did not make compensation and work adjustments and if they, the minimum-wage workers, could make the purchase. The point here is that even worker groups can see the financial benefits of adjusting fringe benefits and work demands in light of a minimum-wage increase.

[20] Even the Employment Policies Institute study cited above (Macpherson, "The Effects of the 1999-2000 Washington Minimum-Wage Increase"), which is likely to contain estimates of the employment losses that are on the high side of the expected range, shows a reduction in Washington's total employment (2.7 million workers) of less than three tenths of one percent for a proposed 26 percent increase in the state's minimum wage. However, it can be noted that if Washington has the average percentage of minimum

This line of argument can also help us understand why workers who retain their jobs are unlikely to be any better off. They get more money, but they also get fewer fringes and have to work harder for their pay. We know the covered workers who retain their jobs will be worse off, at least marginally so, because the only reason an employer intent on making as much profit as possible would offer the fringes and reduced work in the first place is that the workers valued the fringes and lax work demands more highly than they valued the money wages that they had to give up in order to get the fringes or lax work demands. Further, profit-maximizing employers aren't about to offer workers anything that's costly unless they get something in return, like greater output per hour or a lower wage bill.

If a firm offers costly benefits that do not lower wages or fail to offer benefits that could lower wages, then that firm should be subject to takeover. Some savvy investor can be expected to buy the firm, change its benefit policies, lower wages by more than the rise in other costs rise, improving the firm's profitability in the process, and then sell the firm for a higher price.

Make no mistake about it, profit-maximizing firms do not "give" fringes to their workers; they require their workers to pay for the fringes through wage-rate reductions. The wage rate reductions can be expected because, if workers value the fringes, the supply of workers will go up, forcing the money wage rate down.

It follows that competitive market pressures will force firms to do what is right by their bottom lines and their workers. This means that when the minimum wage is raised, the value of the resulting lost fringes and reduced work demands to the workers will be greater than the value of the additional money income.

Put another way, the workers who retain their jobs are made worse off (perhaps, marginally so) in spite of the money-wage increase. Employment in low-skill jobs may go down (albeit ever so slightly) in the face of minimum-wage increases not so much because the employers don't want to offer the jobs (as traditionally argued), but because not as many workers want the minimum-wage jobs that are offered.[21]

Available Empirical Evidence

Have the expected effects been seen in empirical studies? The most compelling evidence is captured in the many studies already cited that indicate that job losses from a minimum-wage increase tend to be small, even within the worker groups are most likely to be adversely affected. However, there have been other studies over the past two

---

wage workers, 8.8 percent, then the EPI study suggests that each 10 percent increase in the minimum wage lowers the employment of covered workers by, at most, 1.2 percent.

[21] Granted, not all low skill workers have many fringe benefits that can be taken away, and some minimum wage workers may be working very hard. The argument that is being developed suggests that the negative employment effects of a minimum wage increase will be concentrated among this group of particularly disadvantaged workers.

decades that have attempted to assess directly the impact of minimum-wage increases on fringes and work demands, as well as the overall value of jobs.

- Writing in the American Economic Review, Masanori Hashimoto found that under the 1967 minimum-wage hike, workers gained 32 cents in money income but lost 41 cents per hour in training -- a net loss of 9 cents an hour in full-income compensation.[22]
- Linda Leighton and Jacob Mincer concluded that increases in the minimum wage reduce on-the-job training -- and, as a result, dampen growth in the real long-run income of covered workers. [23]
- Walter Wessels found that the minimum wage caused retail establishments in New York to increase work demands. In response to a minimum-wage increase, only 714 of the surveyed stores cut back store hours, but 4827 stores reduced the number of workers and/or their employees' hours worked. Thus, in most stores, fewer workers were given fewer hours to do the same work as before.[24]
- The research of Belton Fleisher, [25] William Alpert,[26] and L.F. Dunn[27] shows that minimum-wage increases lead to large reductions in fringe benefits and to worsening of working conditions.

If the minimum wage does not cause employers to make substantial reductions in nonmoney benefits, then increases in the minimum wage should cause (1) an increase in the labor-force participation rates of covered workers (because workers would be moving up their supply-of-labor curves), (2) a reduction in the rate at which covered workers quit their jobs (because their jobs would then be more attractive), and (3) a significant increase in prices of production processes heavily dependent on covered minimum-wage workers. However, Wessels found little empirical support for such conclusions drawn from conventional theory. Indeed, in general, he found that minimum-wage increases had the exact opposite effect: (1) participation rates went down, (2) quit rates went up, and (3) prices did not rise appreciably -- findings consistent only with the view that

---

[22]Masanori Hashimoto, "Minimum Wage Effect on Training to the Job," *American Economic Review*, vol. 70 (December 1982), pp. 1070-87.

[23]Linda Leighton and Jacob Mincer, "Effects of Minimum Wage on Human Capital Formation," in *The Economics of Legal Minimum Wages*," ed. Simon Rothenberg (Washington, D.C.: American Enterprise Institute, 1981).

[24]Walter J. Wessels, "Minimum Wages: Are Workers Really Better Off?" (Paper prepared for presentation at a conference on minimum wages, Washington, D.C., National Chamber Foundation, July 29, 1987). For more details, see Walter J. Wessels, Minimum Wages, Fringe Benefits, and Working Conditions (Washington, D.C.: American Enterprise Institute, 1980).

[25]Belton M. Fleisher, *Minimum Wage Regulation in Retail Trade* (Washington, D.C.: American Enterprise Institute, 1981).

[26] William T. Alpert, The Minimum Wage in the Restaurant Industry (New York: Praeger, 1986).

[27] L.F. Dunn, "Nonpecuniary Job Preferences and Welfare Losses among Migrant Agriculture Workers," *American Journal of Agriculture Economics* 67 (May 1985), pp. 257-65.

minimum-wage increases make workers worse off.[28] With regard to quit rates, Wessels writes,

> I could find no industry which had a significant decrease in their quit rates. Two industries had a significant increase in their quit rates.... These results are only consistent with a lower full compensation. I also found that quit rates went up more in those industries with the average lowest wages, the more full compensation is reduced. I also found that in the long run, several industries experienced a significantly large increase in the quit rate: a result only possible if minimum wages reduce full compensation.[29]

Seen from this perspective, Herbert's cited figures on the added income received by 10 million workers are grossly misleading because the figures suggest that the affected workers are "better off," which is not likely to be the case, given their loss of fringe benefits and increased work demands. The fact that the Card and Krueger studies also found, supposedly, no loss of jobs suggests that the market may have forced non-wage adjustments on the fast food restaurants studied.

Granted, economists might speculate, as they have, that the job reductions have been small because the low-skill labor market exhibits a "low elasticity of demand" (or low responsiveness among employers to a wage hike), but such an explanation is hardly compelling. The demand elasticity for anything, including labor, is related to the number of substitutes the good (or labor) has: the greater the number of substitutes, the greater the ability of buyers (employers) to move away from the good (labor) when the price (wage rate) is raised, and hence, the greater the responsiveness of buyers (employers), or elasticity of demand. The problem with the explanation is that there is no labor group that has more substitutes than low-skill (minimum-wage) workers, especially now that firms have so much flexibility to automate jobs out of existence or to replace domestic workers with foreign workers by way of imports. The elasticity of demand for low-skill labor must be relatively high. Hence, the relatively small decline in the number of low-skill workers in response to a minimum-wage hike points to a conclusion central to this: the mandated wage hike is likely offset in large measure by other adjustments in the affected workers' compensation package.

Minimum-Wage Consequences over Time

This line of argument does not lead to the conclusion that minimum-wage increases of given amounts should always have the exact employment effect no matter when they are legislated. Looking back at Figure 4.7, we might reason that as the real minimum wage rose between 1938 and 1968, employers did what they were pressed to do to moderate their labor cost increases: take away progressively more fringe benefits and add progressively more work demands (compared to what they would have done). Hence, as time went by, we might expect the employment effects of a given minimum wage

[28] Wessels, "Minimum Wages: Are Workers Really Better Off?"
[29] Ibid., p. 13.

increase to go up as 1968 was approached. As time passed, there simply were fewer ways for employers to adjust to the wage hike.

However, as the minimum wage has fallen irregularly since 1968, we might expect employers to respond by gradually adding back more fringe benefits and relaxing their work demands (a trend that has likely been accelerated with growing tightness in labor markets in the late 1990s). The result should be that in the 1990s, employers should have had more ways to adjust to a minimum-wage hike than they had in, say, the late 1960s. As a consequence, we should not be surprised that Card and Krueger found little or no employment effect in the early 1990s, whereas other studies in the 1960s found larger effects.[30] We should not be surprised if future studies of the impact of any 1999 increase in the minimum wage show similarly negligible negative employment effects.

* * * * *

Members of Congress and the president need to recognize a simple fact of modern economics: You can't fool the market as much as imagined, at least not all the time. Politicians simply do not have as much power to manipulate markets as they may think they have. Markets can be expected to outsmart the smartest of politicians in the next round of minimum-wage hikes. We can anticipate that, once again, the chosen increase in the minimum wage will have minimum employment consequences for two reasons: First, members of Congress will choose a fairly small increase in the minimum wage because of political groups working against the proposed minimum-wage bill. Second, market forces will largely neutralize the potential negative employment effects of whatever wage increase is legislated.

Concluding Comments

The market system can perform the very valuable service of rationing scarce resources among those who want them. It alleviates the congestion that develops when resources, goods, and services are rationed by other means. Markets, however, are not always permitted to operate unobstructed. Government has objectives of its own, objectives that are determined collectively rather than individually. We have seen how government can use its power to tax, to raise government revenues, or reallocate market demand.

---

[30] The implication of the theory that a minimum-wage hike will have a greater impact on employment when the minimum wage is high, compared to when it is low, has not been rigorously tested to date. However, it is interesting to note that through the 1950s 1960s, and early 1970s, the editors at the New York Times were staunchly for increases in the minimum wage, mainly because the evidence on the negative employment effect was not strong, to say the least. However, as the evidence in the 1960s mounted that minimum-wage hikes had a negative employment effect, especially among minority teenagers, the editors began to shift their editorial stance. By the mid-1980s, they came out in favor of a minimum wage of "$0.00." They have since shifted their editorial stance back to support for minimum-wage hikes, mainly because the negative employment effects have been shown to be nil in recent studies. See Richard B, Mckenzie, Times Change: *The Minimum Wage and the New York Times* (San Francisco: Pacific Research Institute, 1994).

Government power can also be used to eliminate externalities or reduce monopoly power. Whether the use of such controls is considered good or bad depends to a significant extent on one's personal values and circumstances. In a free market system, price controls and consumer protection will always be controversial.

In the case of minimum wage hikes, it appears that policy makers and economists alike have failed to grasp an important lesson: The hikes do not destroy competition, only redirect its force. They also give managers an incentives to find ways of reducing their impact on employment – and the net benefits of the hikes to the workers.

Review Questions

1. Is a tax on margarine efficient in the economic sense of the term? Why would margarine producers prefer to have an excise tax imposed on both butter and margarine? Would such a tax be more or less efficient than a tax on margarine alone?
2. If punishment for crime is a kind of tax on those who engage in illegal activity, what effect would the legalization of marijuana have on its supply and demand? What would happen to the market price? The quantity sold? Illustrate with supply and demand curves.
3. If in a competitive market, prices are held below market equilibrium by government controls, what will be the effect on output? How might managers be expected to react to the laws?
4. Why might some managers want price controls? Why don't they get together and control prices themselves (if it were legal)?
5. How would price controls affect a firm's incentive to innovate? Explain.
6. "If prices are controlled in only one competitive industry, the resulting shortage will be greater than if prices were controlled in all industries." Do you agree? Explain.
7. "Price controls can be more effective in the short run than in the long run." Explain.
8. Why would some firms want the minimum wage to be increased? Why would some managers who believe that workers "deserve" higher wages cut fringe benefits or increase worker demands in response to a hike in the minimum wage?

READING: **Water Rights and Water Markets**

Terry L. Anderson, Montana State University

Mark Twain wrote, "Whiskey is for drinkin'—water is for fightin'." In the American West, water has always been a matter of survival. It was the cause of many frontier skirmishes, and it may provoke conflict again. Newsweek warned recently that "drought, waste, and pollution threaten a water shortage whose impact may rival the energy crisis." And former Secretary of the Interior James Watt said, "The energy crisis will seem like a Sunday picnic when compared to the water crisis."

Unless Americans change their ways, a water crisis is inevitable. In economic terms, the quantity of water demanded is greater than the quantity available, and there is little time to adjust either amount. The

reason for the imbalance is that the government has been keeping prices below market-clearing levels. In most places in the United States, water is cheaper than dirt. Nowhere in the nation do water prices reflect the true scarcity of the resource.

In Southern California, for example, water is in short supply. Yet Los Angeles residents pay only 0.60 per thousand gallons—a quantity that costs the residents of Frankfurt, Germany, $2.80. It is not surprising, therefore, that each person in the United States consumers an average of 180 gallons a day, compared with 37 gallons in Germany. Water prices are actually lowest in the arid Southwest, where residents of El Paso and Albuquerque pay $0.53 and $0.59, respectively, per thousand gallons, compared with $1.78 in Philadelphia. In many U.S. cities the real price of water has fallen in recent decades, despite the threat of shortages.

Agricultural users, who consume over 80 percent of the water in western states, enjoy extremely low prices. Throughout the nation the price of irrigation water ranges from about $0.009 to $0.09 per thousand gallons. In 1981, the average price of covering one acre of land in California's Central Valley with one foot of water was $5.00, or less than $0.02 per thousand gallons. Supplying that amount of water cost the government as much as $325. According to a 1980 study by the Department of the Interior, government subsidies covered between 57 and 97 percent of the cost of water projects.

Pricing water at market rates could help to solve the water crisis. Water consumption—whether for industrial, municipal, or agricultural use—is highly responsive to price changes. For example, the quantity of water used in industrial processes varies considerably around the world, depending on prices. Where water is expensive, electric power is produced using as little 1.32 gallons per kilowatt-hour. Where water is cheap, production requires as many as 170 gallons per kilowatt-hour. One study of urban water consumption showed that a 10 percent increase in the price of water decreased the quantity of water demanded about 4 to 13 percent.

Pricing water more realistically will require changes in the laws governing water use, as well as the creation of an effective water market. Like any market, a water market will depend on well-defined, well-enforced property rights. If water rights are secure and people can trace them, prices will quickly come to reflect the true scarcity of the resource. During the late nineteenth century, such a system evolved in the American West. Rights were defined on a first-come, first served basis, and institutions arose through which owners of rights could seek out the highest and best use of the resource. The system offered incentives that encouraged some people to deliver water wherever it was demanded. Thousands of miles of ditches were constructed, and millions of acres blossomed, as a result of entrepreneurial efforts to deliver water. Over time, however, legislators, bureaucrats, and judges have tinkered with the system. Legal restrictions now limit the transfer of water, and its use is determined by politicians, not by the market.

One place where a water market might encourage more efficient water use is the Imperial Irrigation District (IID) in Southern California. The IID receives its water from the U.S. Bureau of Reclamation, at subsidized rates. Its water could be conserved if ditches were lined, wastewater recovered, and the timing of irrigation changed. All those measures would be costly to farmers, however. And at present low prices, farmers have little incentive to invest in conservation. Recently, the Municipal Water District (MWD) of Southern California, thwarted in its effort to obtain water from Northern California, has begun negotiating for water from the IID. The MWD is willing to fund improvements in farmers' irrigation systems in return for the water those improvements would save. If such a trade could be accomplished, everyone would be better off.

CHAPTER 5

The Logic of Group Behavior In Business and Elsewhere

Men journey together with a view to particular advantage and by way of providing some particular thing needed for the purpose of life, and similarly the political association seems to have come together originally. . . for the sake of the general *advantage it brings.*

Aristotle[1]

Unless the number of individuals in a group is quite small, or unless there is coercion, . . .rational, self-interested individuals will not act to achieve their common or group interest. In other words, even if all. . . would gain if, as a group, they acted to achieve their common interest or objective, they will still not voluntarily act to achieve that common or group interest.

Mancur Olson[2]

In earlier chapters, we introduced the usefulness of markets. However, as is evident inside firms, not all human interactions are through "markets." People often act cooperatively in groups or, as the case may be, in "firms." In this chapter our central purpose is to explore how and under what conditions people can organize their behavior into voluntary cooperative associations (groups and firms) in which all work together for the attainment of some common objective, say, greater environmental cleanliness, the development of a "club atmosphere," or the maximization of firm profits. The focus of our attention is on the viability of groups like families, cliques, communes, clubs, unions, and professional associations and societies, as well as firms, in which individual participation is voluntary to cohere and pursue the common interests of the members.

We consider two dominant and conflicting theories of group behavior. They are "the common interest theory" and "the economic theory" of group behavior. The former is based on the proposition that a group is an organic whole" identified by the "common interest" shared by its individual members. Its basic thesis is that all groups, even very large ones, are organized to pursue the common interest of the group members. Taking this theory one step further, it implies that if people share a common interest, they will organize themselves into a group and voluntarily pursue their shared interest.

According to the economic theory of group behavior, the group is a collection of independently motivated individuals who organize voluntarily to pursue their common interest only in small groups, like families or clubs. In large groups the common interest

---

[1] Aristotle, *Ethics*, vol. 8, no. 9, p. 1160*a*.
[2] Mancur Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Cambridge, Mass.: Harvard University Press, 1971), p. 2

is very often ineffective in motivating group behavior. The logic of this theory seems perverse; but, as we will see in later chapters, it is the basis for almost all economic discussion of markets and explains why many policy proponents argue governments must be delegated coercive powers to collect taxes and to pursue the "public interest." It also helps explain why firms are organized the way they are and why managers manage the way they do. This is, therefore, one of the pivotal chapters in this book.

However, keep in mind that groups are not the only means by which people's interpersonal or social behavior is organized in society. Economics is basically a study of comparative social systems, an examination of how the different ways of organizing interpersonal behavior can be fitted together in different combinations. We call these means of organizing people's behavior "social organizers" and mention four of them here: markets which involve exchanges of goods and services, government coercion, violence, and voluntary groups. On the surface, violence may not appear to be a bona fide alternative, but we are forced to mention it because of the use made of it throughout the world. The behavior of street-gang members, for example, with respect to people totally unassociated with them, is largely based upon either the existence or the threat of violence. The Cold War was a tenuous truce founded to a sizable degree on the threat of a nuclear holocaust. The persistent violence in the streets of Northern Ireland during the 1960s and 1970s will for many years have a profound influence on what the people of that country can hope to accomplish. Many examples can be cited which illustrate the spread of terrorist activities and the threat they represent to the fabric of social order which has been built on the basis of other social organizers. Aside from what we have already said with regard to anarchy, we will have little to say about violence as a social organizer. This does not lessen the importance, which we attribute to violence; it simply reflects the fact that economists have only recently turned their attention to the subject and much remains to be done in the way of theory construction.[3]

The question of how you appraise the roles the various social organizers should play in social order appears to be wrapped in one's personal ideology or value system—that is, there appears to be no room for positive analysis. Indeed, what we as individuals want the system to accomplish is surely a factor in how each of us evaluates potential social organizers. Personal values will affect our attitude as to whether or not a given social organizer should be used and, if used, how extensively. The avowed Marxist has very harsh opinions of the market system. But perhaps just as important in our appraisal is what we know about the *relative* effectiveness—the advantages and limitations—of the potential means for ordering behavior. If, for example, we have only a rudimentary understanding of how the market works and fail to appreciate with sufficient clarity the limitations of cooperative efforts, we may naturally place greater reliance on voluntary

[3] For example of economists' initial probes into the area of malevolence and violence, see Kenneth E. Boulding, *The Economy of Love and Fear* (Belmont, Calif.: Wadsworth Publishing Company, Inc., 1973), and Gordon Tullock, *The Social Dilemma: The Economics of War and Revolution* (Blacksburg, Va.: University Publications, 1974). Only those who wish to be challenged will find these books useful.

cooperation than we would otherwise. We, therefore, in this chapter highlight the limitations of voluntary groups as a social organizer in order that we may appreciate why markets are not only beneficial but also necessary in organizing a society of heterogeneous individuals.

Common Interest Theory of Group Behavior

There are almost as many theories of group behavior as there are group theorists. However, categorizing theories according to dominant themes or characteristics is sensible in light of our limited space.

All theories of group behavior begin by recognizing the multiplicity of forces, which affect group members and, therefore, groups. This is especially true of what we term the *common interest theory*. Many present-day sociologists, political scientists, and psychologists generally share this point of view, which has been prominent at least since Aristotle. The determinants of group behavior most often singled out are the "leadership quality" of specific group members and the need felt among group members for "affiliation," "security," "recognition," "social status," or money. Groups like clubs or unions form so that members can achieve or satisfy a want that they could not satisfy as efficiently through individual action. All these considerations are instrumental in affecting "group cohesion," which, in turn, affects the "strength" of the group and its ability to compete with other groups for the same objectives. From the perspective of this theory, when people join firms, they accept the firm's objective and pursue it because everyone else wants the same thing, leading to self-enforcing group cohesion.

The common interest theory views the "group" as an *organic whole*, much like an individual, as opposed to a collection of individuals whose separate actions appear to be "group action." According to the theory, the group has a life of its own which is to a degree independent of the individuals who comprise it. Herbert Spencer, a nineteenth-century sociologist, often described the group as a "social organism" or as a "superorganic" entity.[4] Karl Marx wrote of the "class struggle" which will bring down "bourgeois capitalism" and of the proletariat" which will, in its place, erect the communist society. And it was probably the social-organism view of groups that Aristotle had in mind when he wrote, "Man is by nature a political animal."[5]

Two major reasons are given for viewing groups as a social organism. First, a group consists of a mass of interdependencies, which connect the individuals in the group. Without the interdependencies, there would be only isolated individuals, and the term *group* would have no meaning. Individuals are like the nodes of a spider web. The

---

[4] Spencer was actually somewhat ambivalent on the subject; at times he also wrote of groups as a composite of individuals. This aspect of his writing reflected the influence David Hume and Adam Smith had on his thinking. See Herbert Spencer, Principles of Sociology (London: Williams and Norgate, Ltd., 1896).

[5] Aristotle, Politics, Book II.

spider web is constructed on these nodes, and the movements in one part of the web can be transmitted to all other parts. Much like the process of synergism in biology,[6] the actions of individuals within a group combine to form a force that is greater than the sum of the forces generated by individuals isolated from one another. The group must, so the argument goes, be thought of as more than the sum total of individuals. This argument is often used to arouse support for a labor union. Union leaders argue that the union can get higher wage increases for all workers can obtain acting independently of one another. The reason is that union leaders efficiently coordinate the efforts of all. Environmental groups make essentially the same argument: With well-placed lobbyists, the environmental group can have a greater political impact than can all the individuals they represent writing independent letters to their representatives at different times.

Second, groups tend to emerge because they satisfy some interest shared by all the group's members. Because all share this "common interest," individuals have an incentive to work with others to pursue that interest, sharing the costs as they work together. Aristotle wrote, "Men journey together with a view to particular advantage,"[7] and Arthur Bentley said, "There is no group without its interest.. . .The group and the interest are not separate. . . . If we try to take the group without the interest, we simply have nothing at all."[8]

Having observed that a common interest can be shared by all of a group's member, the adherents of this theory of group behavior argue that a group can with slight modification, be treated as an individual. The primary modification is the relative tightness or looseness of the ties that bind the group members together. This usually makes group action and reaction less decisive and precise than that of individuals, but the difference between a group and an individual is still a matter of degree, not kind. For instance, the difficulty of passing information about group goals from person to person can make the group's response to new information somewhat sluggish. Nevertheless, a group can be assumed to maximize the attainment of its common objective. Furthermore, the implicit assumption is made that this will be true of large as well as small groups. It is on this deduction that Mancur Olson and many economists take issue with this analysis of group behavior.

The Economic Theory of Group Behavior

Mancur Olson, on whose work this section rests, agrees that the "common interest" can be influential and is very important in motivating the behavior of members of small groups. However, he, like so many other economists, insists that a group must be looked upon as a composite of individuals as opposed to an anthropomorphic whole, that the common interest, which can be so effective in motivating members of small groups, can be impotent in motivating members of large groups: "Unless there is coercion in large

---

[6] This is the process whereby two or more substances (gases or pollutants) come together, and combined can have a greater effect than the sum of the effects of each individual taken separately.
[7] Aristotle, *Ethics*, vol. 8, no. 1, p. 1160a.
[8] Bentley, in Peter Odegard (ed.) *Process of Government* (Cambridge: The Belknap Press, Harvard University Press, 1967), pp. 211-213.

groups. . . ., *rational self-interested individuals will not act to achieve their common or group interest*." Furthermore, he contends, "These points hold true when there is unanimous agreement in a group about the common goal and the methods of achieving it."[9] To understand this theory, we will first examine the propositions upon which it is founded and then analyze some qualifications.

Basic Propositions

Using economic analysis, people are assumed to be as rational in their decision to join a group as they are toward doing anything else; they will join a group if the benefits of doing so are greater than the costs they must bear. As explained earlier, these costs and benefits, like all others relevant to any other act, must be discounted by the probability that the costs and benefits will be realized.[10] There are several direct, private benefits to belonging to groups, such as companionship, security, recognition, and social status. A person may also belong to a group for no other reason than to receive mail from it and, in that small way, to feel important. A group may serve as an outlet for our altruistic or charitable feelings. If by "common interest" we mean a collection of these types of *private benefits*, it is easy to see how they can motivate group behavior. Entrepreneurs can emerge to "sell" these types of private benefits as they do in the case of private golf clubs or Weight-watchers. The group action will be then, essentially, a market phenomenon—that is, a problem in simple exchange.

However, the central concern of this theory is a "common interest" which is separate and detached from these types of *private* benefits. The concern is with public benefits that transcend the entire group, which cannot be provided by the market, and which may be obtained only by some form of collective action. That is, a group of people must band together to change things from what they otherwise would be. Examples include the common interest of consumers in general to obtain better, safer products *than the market would provide without collective action*; the interest of labor unions is to secure higher wages and better fringe benefits than could be obtained by the independent actions of laborers; the interest of students is to have better instruction; the interest of faculties is to educate quality graduates. These are examples of the common interest being a public good. (As you will recall, a public good was defined as a good—or service—the benefits of which are shared by all members of the relevant group if the good is provided or consumed by anyone.)

---

[9] Olson, *Logic of Collective Action*, p. 2. A number of economists were moving toward the development of Olson's line of analysis, but the force and clarity of Olson's presentation of his view of group behavior make his book an important reference work.

[10] This type of cost and benefit analysis has been explicit, if not implicit, in much of the writing of those in support of the "common interest theory of groups" explained above. There would be little reason for talking about a "common interest" if it did not have something to do with benefits of group participation. See, for example, Dorwin Cartwright, "The Nature of Group Cohesiveness," in Dorwin Cartwright and Alvin Zander, eds., Group Dynamics: Research and Theory, 3d ed. (New York: Harper & Row, Publishers, 1968), pp. 91-109.

Small Groups

Small groups are not without their problems in pursuing the "common interest" of their members. They have a problem of becoming organized, holding together, and ensuring that everyone contributes his part to the group's common interest. This point was illustrated earlier in terms of Fred and Harry's problems of setting up a social contract, and it can be understood in terms of all those little things which we can do with friends and neighbors but which will go undone because of the problems associated with having two or three people come together for the "common good." For example, it may be in the common interest of three neighbors for all to rid their yards of dandelions. If one person does it, and the other two do not, the person who removes the dandelions may find his yard full of them the next year because of seeds doing from the other two yards. Why do we so often find such a small number of neighbors failing to join together to do something like eradicating dandelions?

We can address this question with the use of the public goods demand curve developed earlier. The common interest is dandelion eradication; and two neighbors, Fred and Harry, again, have a demand for this public good.[11] There is no particular reason for us to assume that Fred and Harry have identical demands for this particular public good; consequently, we have drawn Harry's demand for eliminating dandelions in Figure 5.1 greater than Fred's demand.

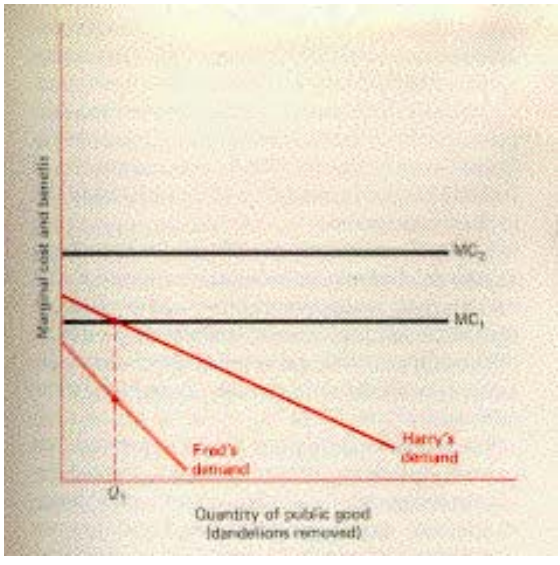


Figure 5.1 The Problem of Getting Collective Action

If the marginal cost curve is MC_2, the marginal cost of eliminating even the first dandelion will be too high to take any action at dandelion eradication. However, if the cost were lower, MC_1 instead of MC_2, Harry would be willing to eliminate as many as Q_1 dandelions. Fred would still do nothing.

[11] We realize the imperfections of this example of a public good; much of the benefit of each person's action is private. Only a portion of one neighbor's dandelions may actually affect other people's yards. The example, however, is a reasonably good one for our purpose.

If we assume, for simplicity only, that the marginal cost of eliminating dandelions is constant, the marginal cost curve will be horizontal. Whether or not either Fred or Harry will individually do anything about his dandelions depends, given their demands, upon the position of the marginal cost curve. If, for example, it is positioned as MC_2 in Figure 5.1, neither Fred nor Harry will be motivated individually to do any thing about the problem. The marginal cost of eliminating the first dandelion is greater than the benefits that even Harry, who has the greater demand, received from it. Notice that the marginal cost curve (MC_2) does not intersect either of the demand curves in the graph, meaning that the optimum level of activity for both, *on an individual basis*, is zero. On the other hand, if the marginal cost curve is at MC_1, Fred will still be unwilling to do anything, but Harry will be willing to eliminate, on his own, up to Q_1 dandelions. Fred, however, will benefit from Harry's actions; he will have fewer dandelions in his own yard; he can "free-ride" because of Harry's high demand for dandelion eradication.

Still, the quantity of dandelions eradicated may not be what is socially optimal. Consider Figure 5.2. In that figure we have constructed Fred and Harry's joint, or public good, demand curve. Their *collective* demand curve is obtained by vertically summing the demands of the individuals. Under individual action, Q_1 dandelions are eradicated by Harry. However, the value which Fred and Harry collectively place on the elimination of additional dandelions is greater than the marginal cost. For example, the marginal value of the Q_1th unit to both Fred and Harry combined is MB_1; the marginal cost, MC_1. They can gain by eliminating that dandelion and all others up to Q_3. This is the point where the marginal cost curve and the public good demand curve intersect. By sharing the cost of eliminating the weeds, they can move to Q_3. Harry will not move to that point if he has to pay the full cost for each unit, MC_1, but he will move beyond Q_1 if he can get Fred to take over part of the cost. How they share the cost must, because of the complications involved, be reserved for a later discussion; we need only point out here that there is no reason to believe that an equal sharing of the cost will be the outcome.

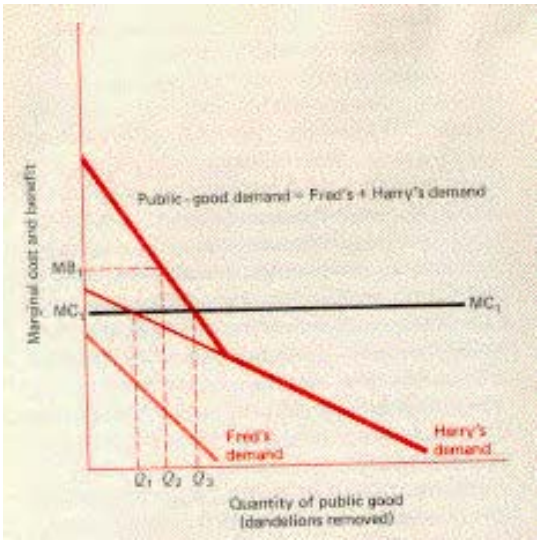


Figure 5.2 Efficient Provision of Collective Goods

The public goods demand curve, which is the darker curve in the figure, is derived by vertically adding the demands of Fred and Harry. Given a marginal cost represented by *MC1*, the optimum quantity of dandelions removed is *Q3*.

Even though Fred and Harry may not ever agree to work out their common problem (or interest) cooperatively, there are several conditions that may lead them to do so. In a small group there is personal contact. Everyone knows everyone else. What benefits or costs there may be from an individual's action are spread over just a few people and, therefore, the effect felt by any one person can be significant. (Fred knows there is a reasonably high probability that what he does to eliminate dandelions from the border of his property affects Harry's welfare.) If the individual providing the public good is concerned about the welfare of those within his group and receives personal satisfaction from knowing that he has in some way helped them, he has an incentive to contribute to the common good; and we *emphasize that before the common good can be realized, individuals must have some motivation for contributing toward it*. Furthermore, "free-riders" are easily detected in a small group. (Harry can tell with relative ease when Fred is not working on, or has not worked on, the dandelions in his yard.) If one person tries to let the others shoulder his share, the absence of his contribution will probably be detected. Others can then bring social pressure to bear to force him to live up to his end of the bargain. The enforcement costs are low because the group is small. There are many ways to let a neighbor know you are displeased with some aspect of his behavior.

Finally, in small groups an individual shirking responsibilities can be excluded from the group if he does not contribute to the common interest and joins the group merely to free ride on the efforts of others. In larger groups, like nations, exclusion is more difficult and, therefore, more unlikely.

The problem of organizing "group behavior" to serve the common interest has been a problem for almost all groups, even the utopian communities that sprang up during the nineteenth century and in the 1960s. Rosebeth Kanter, in her study of successful nineteenth-century utopian communities concluded:

> The primary issue with which a utopian community must cope in order to have the strength and solidarity to endure is its human organization: how people arrange to do the work that the community needs to survive as a group, and how the group in turn manages to satisfy and involve its members over a long period of time. The idealized version of communal life must be meshed with the reality of the work to be done in the community, involving difficult problems of social organization. In utopia, for instance, who takes out the garbage?[12]

Kanter found that the most successful communities minimized the free-rider problems by restricting entry into the community. They restricted entry by requiring potential members to make commitments to the group. A six "commitment mechanism" distinguished the successful from the unsuccessful utopias: (1) sacrifice of habits common to the outside world, such as abstinence from alcohol and tobacco or, in some cases, celibacy; (2) assignment of all worldly goods to the community; (3) adoption of rules which would minimize the disruptive effects of relationships between members and

[12] Rosebeth M. Kanter, *Commitment and Community: Communes and Utopias in Sociological Perspective* (Cambridge, Mass.: Harvard University Press, 1973), p. 64.

nonmembers and which would (through, for example, the wearing of uniforms) distinguish members from nonmembers; (4) collective sharing of all property and all communal work; (5) submission to public confession and criticism; and (6) expressed commitment to an identifiable power structure and tradition. Needless to say, the cost implied in these "commitment mechanisms" would tend to discourage many free riders from joining the society. By identifying the boundaries to societies, these mechanisms made exclusion possible. As Kanter points out, the importance of these commitment mechanisms is illustrated by the fact that their breakdown foreshadowed the end of the community.

Other means of bringing about collective behavior on the part of group members are suggested by the cattlemen's associations formed during the nineteenth century. During the nineteenth century, cattle were allowed to run free over the ranges of the West. The cattlemen had a common interest in ensuring that the ranges were not overstocked and overgrazed and in securing cooperation in rounding up the cattle. To provide for these common interests, cattlemen formed associations which sent out patrols to keep out intruders and which were responsible for the roundups. Any cattleman who failed to contribute his share toward these ends could be excluded from the association, which generally meant that his cattle were excluded from the roundup or were confiscated by the association if they were rounded up.[13]

The family is a small group, which by its very nature is designed to promote the common interest of its members. That common interest may be something called "a happy family life," which is, admittedly, difficult to define. The family does not escape difficulties. At present its validity as a viable institution is being challenged by many sources; however, it does have several redeeming features that we think will cause it to endure, imperfect though it may be, as a basic component of social fabric. Because of the smallness of the group, contributions made toward the common interest of the family can be shared and appreciated directly. Parents usually know when their children are failing to take the interest of the family into account, and children can easily ascertain similar behavior in their parents. Family members are able, at least in most cases, to know personally what others in the group like and dislike; they can set up an interpersonal cost-and-benefit structure among themselves that can guide all members toward the common interest. Most collective decisions are also made with relative ease.[14] However, even with all the advantages of close personal contact, the family as a small group often fails to achieve the common interest. Although all family members may be encouraged to "go their own way" up to a point, some individuals may take this too far. They may fail to contribute their share to the common goal and may cause bitterness and, perhaps, the demise of the family. Given the frequent failure of the family as a viable organization

---

[13] For a very interesting historical investigation of the cattle business during the late nineteenth century, see Rodgers Taylor Dennen, "From Common to Private Property: The Enclosure of the Open Range," Ph.D. dissertation, University of Washington, 1975.

[14] See, for more discussion on the economics of the family, Richard B. McKenzie and Gordon Tullock, "Marriage, Divorce, and the Family," in *The New World of Economics* (Homewood, Ill.: Richard D. Irwin, Inc. 1978), chap. 8

with a common interest,[15] the failure of much larger groups to achieve their expressed common objectives is not difficult to understand.

Large Groups

In a large-group setting, the problems of having individual members contribute toward the development of the common interest are potentially much greater. The direct, personal interface which is present in small groups is usually lacking in larger groups; and, by the nature of large groups and the public good they produce, the benefits generated by any one person are generally spread over a large number of people, so much so that their actions have a significant effect on anyone, even themselves. As a result, they may perceive neither direct benefits in terms of what their behavior does for themselves, personally, nor indirect benefits in terms of what their behavior contributes to the welfare of others.

On the other hand, an individual may be able to detect benefits from his actions, but he must weigh these benefits against the costs he may have to incur to achieve them. For a large group the costs of providing detectable benefits can be substantial—or they can escalate with the size of the group. This is not only because there are more people to be served by the public good,[16] but also because large groups are normally organized to provide public goods that are rather expensive to begin with. Police protection, national defense, and schools are examples of very costly public goods provided by large groups. If all people contribute to the public good, the cost to any one person can be slight; but the question confronting the individual is how much he will have to contribute to make his actions detectable, *given what all the others do*.

In the context of a very large group, suppose there are certain common national objectives to which we can all subscribe, such as a specific charitable program. It is, in other words, in our "common interest" to promote this program. Will people be willing to voluntarily contribute to the federal treasury for the purpose of achieving this goal? Certainly some people will (as Harry does in Figure 5.1 with a marginal cost of MC_2), but many people may not. As they do each April 15 (the deadline for filing tax returns), most will contribute as little income tax as possible. Under a system of voluntary contributions, some people will contribute nothing. A person may reason that although he agrees with the national objective, or common interest, his contribution—that which he can justify—will do little to achieve it. He can also reason that withholding his contribution will have no detectable effect on the scope and effectiveness of the program. (If you or your parents did not pay taxes, would the level of public goods that benefit you

[15] Approximately one-third of all families based on the institution of marriage end in divorce. Many others fail, in terms of the presence of intense hostility, even though there is no legal recognition of that fact.
[16] For a pure public good, the costs, by definition, do not rise with a few additional members. However, most groups provide services that are less than a *pure* public good. Education is an example of an impure public good; all education does not benefit all members of society simultaneously and to the same degree. Under these circumstances, the costs can rise, as we have suggested, with the membership, although by a lower percentage.

be materially affected?) It is for this reason that compulsory taxes are necessary. Olson writes:

> Almost any government is economically beneficial to its citizens, in that the law and order it provides is a prerequisite to all civilized economic activity. But despite the force of patriotism, the appeal of the national ideology, the bond of a common culture, and the indispensability of the system of law and order no major state in modern history has been able to support itself through voluntary dues or contributions. Philanthropic contributions are not even a significant source of revenues for most countries. Taxes, compulsory payments by definition, are needed. Indeed, as the old saying indicates, their necessity is as certain as death itself.[17]

The general tenor of the argument also applies to contributions that go to CARE, a voluntary charitable organization interested mainly in improving the diets of impoverished people around the world. Many of the students reading these pages have been disturbed by scenes of undernourished and malnourished children shown in television commercials for CARE. All those who are disturbed would probably like to see something done for these children. They have had an opportunity to make a contribution, but how many people ever actually contribute so much has a dollar? Needless to say, many do give. They are like Harry in Figure 5.2, who is willing to dig, voluntarily, some of the weeds from his yard. On the other hand, we emphasize the point that a large number of people who have been concerned never make a contribution. (It would be an interesting classroom experiment to see how many students are disturbed by the CARE commercials and how many have ever given to the organization.) There are many reasons for people not giving, and we do not mean to understate the importance of these reasons; we mean only to emphasize that the large-group problem is one significant reason.

True, if all members of a large group make a small contribution toward the common interest, whatever it is, there may be sizable benefits to all within the group. But, again, the problem that must be overcome is the potential lack of *individual* incentives form which he collective behavior must emerge. Through appropriate organization of group members, the common interest *may* be achieved, even if the membership is large. This, however, merely shifts our attention to the problem of developing that organization. The organization of a large group can be construed as a public good, and there are likely to be costs to making the organization workable. This is likely for two reasons: first, there are a large number of people to organize, which means that even if there is no resistance on the part of the people to be organized, there will be costs associated with getting them together or having them work at the same time for the same objectives. Second, some individuals may try to "free-ride" on the efforts of others, which means it will cost more to get people to become members of the group. Further, each free rider implies a greater burden on the active members of the group. If everyone waits for "the other guy to take the initiative," the group may never be organized. It is because of the organization costs

[17] Olson, Logic of Collective Action, p. 13

that students complain so often about the instructional quality of the faculty or some other aspect of university life without doing anything about it. This is also why most people who are disgruntled with the two major political parties do not form a party with those who share their views. The probability of getting sufficient support is frequently very low, which is another way of saying the expected costs are high.

Because an organization may appear to be an obvious way to promote the public good, individuals who try to organize people for that purpose may go through a learning experience before they conclude that it is too costly a venture for them. Even if the organization is successful, the success may be temporary. Eventually, the free-rider problem emerges and the group may fall apart. During the winter of 1973-74, the United States was in the midst of an "energy crisis." Prices of gasoline and other fuels were being held down in spite of the limited imports of fuel coming into the country from the Middle Eeast. Truckers were having a difficult item obtaining adequate supplies of diesel fuel and of passing their higher operating costs through to the buyers of truck services. Independent truckers sensed that it was in their common interest (not the public's, of course,) to halt their deliveries of goods and services and, in that way, put pressure on the authorities to increase rates and to allocate more fuel supplies for the use of truckers. The call for cooperation met with some success; some truckers did terminate operations and some caught headlines by blocking traffic on major highways. However, there were many unwilling to go along with the work stoppage—something that was in their *common* interest. Consequently, the supporters of the work stoppage resorted to violence, and it was the threat of violence, and not the common interest, which kept many truckers off the road. If it had not been for the violence and the initial willingness of state police departments to allow truckers to flaunt the law by stopping traffic, including other truckers, it is very doubtful that the truckers would have had as much success as they did.

Qualifications to the Economic Theory

Obviously, there are many cases in which people acting in what may appear to be rather large groups try to accomplish things that are in the common interest of the membership. The League of Women Voters during the mid-1970s pushed hard for passage of the Equal Rights Amendment. To the Constitution; labor unions work for wage increases; and the American Medial Association does lobby for legislation that is in the common interest of a large number of doctors. Churches, the Blood Mobile, and other charitable groups are able to work fairly effectively for the "public interest," and several of the possible explanations for this observed behavior force us to step outside the scope of the public goods theory.

Why may people work for the "public interest"? First, as Immanuel Kant, an eighteenth century philosopher, said they should, people can place value on the *act* itself as distinguished from the results or consequences of the act. The *act* of making a charitable contribution, which can be broadly defined to include picking up trash in public areas or holding the door for someone with an armful of packages, may have a value in and of itself. This is true whether the effects of the act are detectable to the individual making he charitable contribution or not. The personal satisfaction (or value)

that comes form the act itself is probably the dominant reason why some people do give to CARE. To the extent people behave in this way, the public good theory loses force. Notice, however, that Olson, in formulating his argument, focused on rational economic man as opposed to moral man, envisioned by Kant. We expect that as the group becomes larger, greater effort will be made to instill people with the belief that the *act* itself is important.

Second, the contribution that a person has to make in group settings is often so slight that even though the private benefits are small, the contribution to the common interest is also small and can be a rational policy course. This may explain, for example, student membership in groups like the National Association of Student Teachers. All one has to do in many situations like this one is show up at an occasional meeting and make a small dues payment. Further, the private benefits of being with others at the meetings and finding out what the plans are for the association can be sufficient incentive to motivate limited action that is in the common interest.

Third, all may not equally share the benefits received by group members from promotion of the common interest. One or more persons may receive a sizable portion of the total benefits and, accordingly, be willing to provide the public good, at least up to some limit. Many businessmen are willing to participate in local politics or to support advertising campaigns to promote their community as a recreational area. Although a restaurant owner may believe the entire community will benefit economically from an influx of tourists, he is surely aware that a share of these benefits will accrue to himself. Businessmen may also support such community efforts because of implied threats of being socially ostracized.

Fourth, large organizations can be broken down into smaller groups. Because of the personal contact with the smaller units, the common interest of the unit can be realized. In promoting the interest of the small unit to which they belong, people can promote the common interest of the large group. The League of Women Voters is broken down into small community clubs that promote interests common to other League clubs around the country. The Lions Club collectively promotes programs to prevent blindness and to help the blind; they do this through a highly decentralized organizational structure. Political parties are structured in such a way that the local precinct units "get out the votes." The surest way for a presidential contender to lose an election is to fail to have a "grass-roots" (meaning small-group) organization. Churches are organized into congregations, and each congregation is decentralized further into circles and fellowship groups. Most of the work in the Congress is done in committees and subcommittees. Quiet often a multiplicity of small groups is actually responsible for what may appear to be the activity of a large groups. The decentralization that is so prevalent among voluntary groups tends to support the economic view of groups[18]

[18] Admittedly, other explanations for decentralization can be made, one of which relates to *diseconomies of scale*. That is, the organization just becomes technically less efficient as its size is expanded. The economic theory of groups rests on the motivational aspect of large organizations, rather than on the technical *capabilities* of the organization.

Fifth, large groups may be viable because the group organizers sell their members a service and use the profits from sales to promote projects that are in the common interest of the group. The Sierra Club, which is in the forefront of the environmental movement, is a rather large group that has members in every part of North America. The group receives voluntary contributions from members and nonmembers alike to research and lobby for environmental issues. However, it also sells a number of publications and offers a variety of environmentally related tours for its members. From these activities, it secures substantial resources to promote the common interest of its membership. The American Economics Association has several thousand members. However, most economists do not belong to the AEA for what they can do for it. They join primarily to receive its journal and to be able to tell others that they belong—both, private benefits. (The AEA also provides economists with information on employment opportunities.)

Sixth, the basic argument for any group is that people can accomplish more through groups than they can through independent action. This means that there are potential benefits to be reaped (or, some may say, "skimmed off") by anyone who is willing to bear the cost of developing and maintaining the organization. A business firm is fundamentally a *group* of workers and stockholders interested in producing a good (a public good, to them). They have a common interest in seeing a good produced which will sell. The entrepreneur is essentially a person who organizes a group of people into a production unit; he overcomes all the problems associated with trying to get a large number of people to work in their common interest by providing workers with private benefits -- that is, he pays them for their contribution to the production of the good. The entrepreneur-manager can be viewed as a person who is responsible for reducing any tendency of workers to avoid their responsibilities to the large-group firm. Because it is in their interest to eliminate shirking, the workers may be just as interested as stockholders in having and paying someone to perform this task.[19] An individual worker may be delighted if he is allowed to remain idle while no one else is, but he will want to avoid the risks of all workers shirking. If all shirk, nothing will be sold, the firm will collapse, and workers will lose their wages. We may, therefore, expect that even in communist societies, managers will be paid handsomely (relatively speaking) for the tasks they perform. It is interesting to note that the wage differential between workers and managers is greater in the Soviet Union than it is in the United States.[20]

MANAGER'S CORNER I: **The Value of Tough Bosses**

What does the "logic of group behavior" have to do with the direct interest of MBA students who seek to run businesses and direct the work of others? In a word, "plenty," as we will see throughout the rest of the book. We will show how the "logic" is central to

[19] These points have been made in a much more complete and technical manner by Armenia A. Alton and Harold Demotes, "Production, Information Costs, and Economic Organization," American Economic Review, vol. 62, pp. 777-795, December 1972

[20] Some managers in the Soviet Union are paid less than industrial workers in the United States; however, the ratio of a manager's salary to a worker's salary is typically greater in the Soviet Union.

how competitive markets (and cartels) work and will discuss a multitude of ways to apply the "logic" directly to management problems.

For now, we can stress a maxim that emerges from the economic view of group behavior: Being (or having) a tough boss is tough, but a boss who isn't tough isn't worth much. And because tough bosses are valuable, and lenient bosses are not, there is a reason for believing that existing organizational arrangements serve to impose the discipline on bosses necessary to ensure that they do a good job imposing discipline on the workforce. Competition will press firms to hire tough bosses, and, as we will show in this chapter, the owners of the firm, or their manager-agents, not workers, will tend to the bosses. That is to say, owners or their agents will tend to boss workers, not the other way around, for the simple reason that worker-bosses will not likely survive in competitive markets. Workers may not like tough bosses, but we will explain that, if given the option, workers would choose to hire tough bosses.[21]

Everyone recognizes that firms compete with each other by providing better products at lower prices in a constant effort to capture the consumer dollar. This competition takes place on a number of fronts, including innovative new products, cost cutting production techniques, clever and informative advertising, and the right pricing policy. But a continuing theme of this and other management books is that none of these competitive efforts can be successful unless a firm backs them up with an organizational structure that is competitive -- one that motivates its employees to work diligently and cooperatively. Before addressing the issue of organization, however, let's first examine why workers value tough bosses. Those firms that do the best job in this organizational competition are the most likely to survive and thrive.

The organizational arrangements used by the most successful firms are most likely to be adopted by other firms, because of the force of profit maximization and market competition. So we should expect business firms to be organized in ways that motivate bosses to work diligently at motivating workers to work diligently and at the least cost. We should expect that the choice between workers and owners of capital as to which group will market the better bosses will depend on which group can be expected to press the other to work the most diligently or at the least cost. We have already given away the answer: Owners (or their manager-agents) will tend to boss the workers, a perfectly acceptable outcome for the owners, of course, but also for the workers, which might not be expected. To understand that point, we must first appreciate why workers would want tough bosses.

Take this Job and . . .

Though probably overstated, common wisdom has it that workers do not like their bosses, much less tough bosses. The sentiment expressed in the well-known country song "Take This Job and Shove It" could only be directed at a boss. Bosses are also the butts of much humor. There is the old quip that boss spelled backward is "Double SOB."

---

[21] As we will see, even when workers own the firm and could be their own bosses, they invariably hire a boss, typically a tough one at that.

And there is the story about the fellow who went to the president of a major university and offered his services as a full professor. Noticing that the fellow had no advanced degree, the president informed him that he was unqualified. The fellow then offered his services as an associate professor and received the same response. After offering his services as an assistant professor and hearing that he was still unqualified, the fellow muttered. "I'll be a Son-of-a-Bitch," at which point the president said, "Why didn't you tell me earlier? I'm looking for someone to be dean of the business school."

If it were not for an element of truth contained in them, such jokes would be hopelessly unfunny. Bosses are often unpopular with those they boss. But tough bosses are much like foul tasting medicines are for the sick; you don't like them, but you want them anyway because they are good for you. Workers may not like tough bosses, but they willingly put up with them because tough bosses mean higher productivity, more job security, and better wages.

The productivity of workers is an important factor in determining their wages.[22] More productive workers receive higher wages than less productive workers. Firms would soon go bankrupt if they paid workers more than their productivity indicated they should be paid, but firms would soon lose their workers if they paid them less than their productivity.

Many things, of course, determine how productive workers are. The amount of physical capital they work with, and the amount of experience and education (human capital) the workers bring to their jobs are two extremely important, and commonly discussed, factors in worker productivity. But how well the workers in a firm work together as a team is also important (a point that will become more apparent in the "Manager's Corner" on "The Value of Teams" later in this chapter). An individual worker can have all the training, capital and diligence needed to be highly productive, but productivity will suffer unless other workers pull their weight by properly performing their duties. The productivity of each worker is crucially dependent upon the efforts of *all* workers in the vast majority of firms.

So *all* workers are better off if they *all* work conscientiously on their *individual* tasks and as part of a team. In other words, it is collectively rational for everyone to work responsibly. But there is little individual motivation to work hard to promote the collective interest of the group, or firm.[23]

While each worker wants other workers to work hard to maintain the general productivity of the firm, each worker recognizes that her contribution to the general productivity is small. By shirking some responsibilities, she receives all of the benefits from the extra leisure but suffers from only a very small portion of the resulting productivity loss, which is spread over everyone in the firm. She suffers, of course, from some of the productivity loss when other workers choose to loaf on the job, but she

---

[22] It is also true, as we will see in a later chapter, that how wages are paid can be an important factor in determining how productive workers are.

[23] This line of analysis has been developed at length by Mancur Olson, The Logic of Collective Action: Public Goods and the Theory of Groups (Cambridge, Mass.: Harvard University Press, 1965).

knows that the decisions others make are independent of whether she shirks or not. And if everyone else shirks, little good will result for her, or for the firm, from diligent effort on her part. So no matter what she believes other workers will do, the rational thing for her to do is to capture the private benefits from shirking at practically every opportunity. With all other workers facing the same incentives, the strong tendency is for shirking on the job to reduce the productivity, and the wages, of all workers in the firm, and quite possibly to threaten their jobs by threatening the firm's viability.

The situation just described is another example of the general problem of the logic of group behavior, or more precisely a form of the **prisoners' dilemma** that is endemic to that logic. This involves a classic police interrogation technique in which officers separate two suspects, indicating to each that if she confesses, then she will get off with light charges and penalties. Collectively, they might both be better off if neither confesses (which implies that the two suspects work together for their common objective, a lighter sentence), but each can be even better off if she confesses while her cohort doesn't. More formally, a prisoners' dilemma is a situation in which each individual is better off by acting independently of other parties in the group, no matter what the other parties do, but all parties in the group are better off by working together.

Consider a slightly different form of the prisoner's dilemma that is described in the matrix in Table 5.1, which shows the payoff to Jane for different combinations of shirking on her part and shirking on the part of her fellow workers.[24] No matter what Jane believes others will do, the biggest payoff to her (in terms of the value of her expected financial compensation and leisure time) comes from shirking. Clearly, she hopes everyone else works responsibly so that general labor productivity and the firm's profits are high despite her lack of effort, in which case she receives the highest possible payoff that any one individual can receive of 125.[25] Unfortunately for Jane, all workers face payoff possibilities similar to the ones she faces (and to simplify the discussion, we assume everyone faces the same payoffs). So everyone will shirk which means that everyone will end up with a payoff of 50, which is the lowest possible collective payoff for workers.[26]

Workers are faced with self-destructive incentives when their work environment is described by the shirking version of the prisoners' dilemma (which we have discussed now in several other contexts). It is clearly desirable for workers to extricate themselves from this prisoners' dilemma. They can double their gain. But how?

---

[24] The payoff can be in dollars, utility, or any other unit of measure. The only important consideration is that higher numbers represent higher payoffs. This is in contrast to the original prisoners' dilemma example in which the number in the payoff matrix represented the length of prison sentences, so the higher number represented lower payoffs.

[25] Of course, not everyone can receive this payoff.

[26] Jane would receive a lower payoff of 25 if she were the only one who did not shirk, but because of her effort the collective payoff would be higher than if she did shirk, as her effort would raise the payoff to the shirkers to something slightly higher than 50.

Table 5.1 The Inclination to shirk on the Job

		Other Workers		
		None shirk	**Some shirk**	**All shirk**
Jane	**Don't shirk**	100	75	25
	Shirk	125	100	50

In an abstract sense, the only way to escape this prisoners' dilemma is to somehow alter the payoffs for shirking. More concretely, this requires workers to agree to collectively subject themselves to tough penalties that no one individual would unilaterally be willing to accept. While no one will like being subjected to tough penalties, everyone will be willing to accept the discipline those penalties impose in return for having that discipline applied to everyone else.

The situation here is analogous to many other situations we find ourselves in. For example, consider the problem of controlling pollution that was briefly mentioned in an earlier chapter. While each person would find it convenient to be able to freely pollute the environment, when everyone is free to do so we each lose more from the pollution of others than we gain from our own freedom to pollute. So we accept restrictions on our own polluting behavior in return for having restrictions imposed on the polluting behavior of others. Littering and shirking may not often be thought of as analogous, but they are. One pollutes the outside environment and the other pollutes the work environment.

An even better analogy is that between workers and college students. The "productivity" of a college from the student's perspective depends on its reputation for turning out well-educated graduates with high grade a reliable indication that a student has worked hard and learned a lot. But students are tempted to take courses from professors who let them spend more time at parties than in the library and still give high grades. But if all professors curried favor with their students with lax grading policies, all students would be harmed as the value of their degrees decreased. While students may not like the discipline imposed on them by tough professors, they want tough professors to help them maintain the reputation of their college and the value of their diplomas. (The ideal situation for each student is for the professor to go easy on him or her alone and to be demanding of all other students.[27])

Similarly, workers may not like bosses who carefully monitor their behavior, spot the shirkers and ruthlessly penalize them, but they want such bosses. We mean penalties sufficiently harsh to change the payoffs in Table5.1 and eliminate the prisoners' dilemma. As shown in Table 5.1, the representative worker Jane captures 25 units of benefits from shirking no matter what other workers do. If she had a boss tough enough to impose more than 25 units of suffering, say 35 units, on Jane if she engaged in shirking, her relevant payoff matrix would be transformed into the one shown in Table 5.2. Jane may not like her new boss, but she would cease to find advantages in shirking. And with a

[27] See Dwight Lee, "Why It Pays to Have Tough Profs," The Margin (September/October 1990): 28-29.

tough boss monitoring all workers, and unmercifully penalizing those who dare shirk, Jane will find that she is more than compensated because her fellow workers have also quit shirking. Instead of being in an unproductive firm, surrounded by a bunch of other unproductive workers, each receiving a payoff of 50, she will find herself as part of a hard-working, cooperative team of workers, each receiving a payoff of 100.

The common perception is that bosses hire workers, and in most situations this is what appears to happen. Bosses see benefits that can be realized only by having workers, and so they hire them. But since it is also true that workers see benefits that can be realized only from having a boss, it is reasonable to think of workers hiring a boss, and preferably a tough one.

Table 5.2 Shirking in Large Worker Groups

		Other Workers		
		None shirk	**Some shirk**	**All shirk**
Jane	**Don't shirk**	100	75	25
	Shirk	90	65	15

Actual Tough Bosses

The idea of workers hiring a tough boss is illustrated by an interesting, though probably apocryphal, story of a missionary in 19th century China. Soon after arriving in China, the missionary, who was then full of enthusiasm for doing good, came upon a group of men pulling a heavily loaded barge up a river. Each man was holding on to a rope attached to the barge as he struggled forward against the river's current, while on the barge was a large Chinaman with a long whip with which he lashed the back of anyone who let his rope go slack. Upon seeing this, the missionary experienced a surge of indignation and rushed up to the group of Chinamen to inform them that he would put an end to such outrageous abuse. Instead of being appreciative of the missionary's concern, however, the Chinamen told him to butt out, that they owned the barge, they earned more money the faster they got the cargo up the river, and they had hired the brute with the whip to eliminate the temptation each would otherwise have to slack off.

The missionary story may be doubted, but the point shouldn't be. Even highly skilled and disciplined workers can benefit from having a "boss" help them overcome the shirking that can be motivated by the prisoners' dilemma. Consider the experience related by Gordon E. Moore, a highly regarded scientist and one of the founders of Intel, Inc. Before Intel, Moore and seven other scientists entered a business venture that failed because of what Moore described as "chaos." Because of the inability of the group of scientists to act as an effective team in this initial venture, before embarking on their

next, according to Moore, "the first thing we had to do was to hire our own boss -- essentially hire someone to run the company."[28]

Pointing to stories and actual cases where the workers hire their boss is instructive in emphasizing the importance of tough bosses to workers. But the typical situation finds the boss hiring the workers, not the other way around. We will explain later why this is the case, but we can lay the groundwork for such an explanation by recognizing that our discussion of the advantages of having tough bosses has left an important question unanswered. An important job of bosses is to monitor workers and impose penalties on those who shirk, but how do we make sure that the bosses don't shirk themselves? How can you organize a firm to make sure that bosses are tough?

The work of a boss is not easy or pleasant. It requires serious effort to keep close tabs on a group of workers. It is not always easy to know when a worker is really shirking or just taking a justifiable break. A certain amount of what appears to be shirking at the moment has to be allowed for workers to be fully productive over the long run. There is always some tension between reasonable flexibility and credible predictability in enforcing the rules, and it is difficult to strike the best balance. Too much flexibility can lead to an undisciplined workforce, and too much rigidity can destroy worker morale. Also, quite apart from the difficulty of knowing when to impose tough penalties on a worker is the unpleasantness of doing so. Few people enjoy disciplining those they work with by giving them unsatisfactory progress reports, reducing their pay, or dismissing them. The easiest thing for a boss to do is not to be tough on shirkers. But the boss who is not tough on shirkers is also a shirker.

A boss can also be tempted to form an alliance with a group of workers who provide favors in return for letting them shirk more than other workers. Such a group improves its well being at the expense of the firm's productivity, but most of this cost can be shifted to those outside the alliance.

Of course, you could always have someone whose job it is to monitor the boss and penalize him when he shirks on his responsibility to penalize workers who are shirking. But two problems with this solution immediately come to mind. One, the second boss will be even more removed from workers than the first boss, and so will have an even more difficult time knowing whether the workers are being properly disciplined. Second, and even more important, who is going to monitor the second boss and penalize him or her for shirking? Who is going to monitor the monitor? This approach leads to an infinite regression, which means it leads nowhere. The solution to the problem is the one workers should want by making sure that the boss has some incentive to be tough. The workers should want their bosses to be "incentivized" to remain tough in spite of all the temptations to concede in particular circumstances for particular workers.

---

[28] See Gordon E. Moore, "The Accidental Entrepreneur," Engineering & Science, vol. 62, no. 4 (Summer 1994): 23-30.

The Role of the Residual Claimant

Every good boss understands that he or she has to be more than just "tough." A boss needs to be a good "leader," a good "coach," and a good "nurse maid," as well as many other things. The good boss inspires allegiance to the firm and the commonly shared, corporate goals. Every good boss wants workers to seek the cooperative solutions in the various prisoners' dilemmas that invariably arise in the workplace. Having said that, however, a good boss will invariably be called upon to make some pretty tough decisions, mainly because the boss usually stands astride the interests of the owners above and the workers below. The lesson of this "Manager's Corner" to this point should not be forgotten, "Woe be to the boss who simply seeks to be a nice guy to all claims." But firms must structure themselves so that bosses will *want* to be tough. How can that be done?

In many firms the boss is also the owner. The owner/boss is someone who owns the physical capital (such as the building, the land, the machinery, and the office furniture), provides the raw materials and other supplies used in the business, and hires and supervises the workers necessary to convert those factors of production into goods and services. In return for assuming the responsibility of paying for all of the productive inputs, including labor, the owner earns the right to all of the revenue generated by those inputs.

Economists refer to the owners as residual claimants (a concept first introduced in our discussion of property rights), since they are the ones who claim any residual (commonly referred to as profits) that remains from the sales revenue after all the expenses have been paid. As the boss, the owner is responsible for monitoring the workers to see if each one of them is properly performing his or her job, and for applying the appropriate penalties (or encouragement) if they aren't. By combining the roles of ownership and boss in the same individual, a boss is created who, as a residual claimant, has a powerful incentive to work hard at being a tough boss.

The employees who have the toughest bosses are likely to be those who work for residual claimants. But the residual claimants probably have the toughest boss of all -- themselves. There is a lot of truth to the old saying that when you run your own business, you are the toughest boss you will ever have. Small business owners commonly work long and hard since there is a very direct and immediate connection between their efforts and their income.[29] When they are able to obtain more output from their workers, they increase the residual they are able to claim for themselves. A residual-claimant boss may be uncomfortable disciplining those who work for her, or dismissing someone who is not doing the job, and indeed may choose to ignore some shirking. But in this case the cost of the shirking is concentrated on the boss who allows it, rather than diffused over a large number of people who individually have little control over the shirking and little motivation to do anything about it even if they did. So with a boss who is also a residual

[29] For example, in 1992 wage and salary agricultural workers averaged a 40.6-hour week, while self-employed agricultural workers averaged a 47.1-hour week. See United States Bureau of the Census, Statistical Abstract of the United States: 1993 (113th edition), Washington, DC, 1993: p. 401, table 636.

claimant, there is little danger that shirking on the part of workers will be allowed to get out of hand.

When productive activity is organized by a residual claimant, all resources -- not just labor -- tend to be employed more productively than when those who make the management decisions are not residual claimants. The contrast between government agencies and private firms managed by owner/bosses, or proprietors, is instructive. Examples abound of the panic that seizes the managers of public agencies at the end of the budget year if their agencies have not spent all of the year's appropriations. The managers of public agencies are not claimants to the difference between the value their agency creates and the cost of creating the value. This does not mean that public agencies have no incentive to economize on resources, only that their incentives to do so are impaired by the absence of direct, close-at-hand residual claimants.[30]

If, for example, a public agency managed to perform the same service for a hundred thousand dollars a year less than in previous years, the agency administrator would not benefit by being able to put the savings in her pocket. In fact, she would find herself worse off as she would be in charge of an agency with a smaller budget and therefore one less prestigious in the political pecking order. She would also realize that the money she saved by her diligence would be captured by an over-budgeted agency, enhancing the prestige of its less efficient administrator.

The clever public administrator is one who makes sure every last cent, and more, of the budget is spent by the end of the budget year, regardless of whether it is spent on anything that actually improves productivity. Can you imagine a proprietor of a private firm responding to the news that production costs are less than expected by urging his employees to buy more computers and office furniture, and attend more conferences before the end of the year?[31]

To make the point differently, assume that as a result of your management training you become an expert on maximizing the efficiency of trash pick-up services. In one nearby town the trash is picked up by the municipal sanitation department, financed out of tax revenue, and headed by a public spirited, bureaucratic sanitation professional. In another nearby town the trash is picked up by a private firm, financed by direct consumer charges, and owned by a local businessperson who is proud of her loyal workers and impressive fleet of trash trucks. By applying linear programming techniques

---

[30] Granted, taxpayers could be viewed as the residual claimants to any efficiency improvement resulting from tough managerial decisions in public enterprises, given that efficiency improvement can result in lower tax bills. However, taxpayers have little incentive to closely monitor the activities of public agencies, and, as a matter of fact, do little of it. The reason is simple: Each taxpayer can reason that there is little direct payoff to anyone incurring the costs of monitoring and enforcing greater efficiency in public agencies. [See Gordon Tullock, The Mathematics of Politics (Ann Arbor, Mich.: University of Michigan Press, 1972), especially chap. 7.]

[31] You might expect a manager down in the bowels of a large corporation urging his workers to "waste" money at the end of the year, but not someone who has a substantial stake in his or her own decisions. The single proprietor/residual claimant is someone who has total claim to the net income stream, which implies maximum incentive to minimize waste.

to the routing pattern, you discover that each trash service can continue to provide the same pickup with half the number of trucks and personnel currently being used.

Who is going to be most receptive to your consulting proposal to streamline their trash pickup operation, the bureaucratic manager who never misses an opportunity to tell of his devotion to the taxpaying public, or the proprietor who is devoted to her workers and treasures her trash trucks? Bet on this, the bureaucrat will show you the door as soon as he becomes convinced that your idea really would save a lot of taxpayer dollars by reducing his budget by 50 percent.

On the other hand, the proprietor will hire you as a consultant as soon as she becomes convinced that your ideas will allow her to lay off half of her workers and sell half of her trucks. The manager who is also a residual claimant can be depended on to economize on resources despite his or her other concerns. The manager who is not a residual claimant can be depended on to waste resources despite his or her statements to the contrary.[32]

No matter how cheaply a service is produced, resources have to be employed that could have otherwise been used to produce other things of value. The value of the sacrificed alternative has to be known and taken into account to make sure that the right amount of the service is produced. As a residual claimant, a proprietor not only has a strong motivation to produce a service as cheaply as possible, she also has the information and motivation to increase the output of the service only as long as the additional value generated is greater than the value foregone elsewhere in the economy.

The prices of labor and other productive inputs are the best indicators of the value of those resources in their best alternative uses. So the total wage and input expense of a firm reflects quite well the value sacrificed elsewhere in the economy to manufacture that firm's product. Similarly, the revenue obtained from selling the product is a reasonable reflection of the product's value. So proprietors of businesses receive a constant flow of information on the net value their firm is contributing to the economy, and self-interest motivates a constant effort to produce any given level of output, and produce it in the way that maximizes firms' contributions.

When the one controlling the firm can claim a firm's profits, those profits serve a very useful function in guiding resources into their most valuable uses. If, for example, consumers increase the value they place on musical earrings (if such were ever made) relative to the value they place on other products, the price of musical earrings will increase in response to increased demand, as will the profits of the firms producing them. The increased profit will give the proprietors of these firms the financial ability, and the motivation, to obtain additional inputs to expand output of this dual-purpose fashion accessory of which consumers now want more. Also, some proprietors of firms making other products will now experience declining profits and find advantages in shifting into

---

[32] Much of the motivation for privatizing municipal services comes from the cost reductions that take place when residual claimants are in charge of supplying these services. There is plenty of evidence that privatization does significantly lower the cost, often by 50 percent or more, of basic municipal services such as trash pick-up, fire protection, and school buses. See James T. Bennett and Manual H. Johnson, Better Government at Half the Price (Ottawa, Ill.: Carolina House, 1983).

production of musical earrings. This redirection of labor and other productive resources continues, driving down prices and profits in musical earring production, until the return in this productive activity is no greater than the return in other productive activities. At this point there is no way to further redirect resources to increase the net value they generate.[33]

The incentives created by residual-claimant business arrangements do a reasonable job of lining up the interests of bosses with the interests of their workers, their customers, and the general goal of economic efficiency -- using scarce resources to create as much wealth as possible. This alignment of interests is a crucial factor in getting large numbers of people with diverse objectives and limited concern for the objectives of others to cooperate with one another in ways that promote their general well being. Having the residual claimant direct resources is, understandably, an organizational arrangement that workers should applaud. The residual claimant can be expected to press all workers to work diligently, so that wages, fringes, and job security can be enhanced. Indeed, the workers would be willing to pay the residual claimants to force all workers to apply themselves diligently (which is what they effectively do); both workers and residual claimants can share in the added productivity from added diligence.

Certainly this ability to productively harmonize a diversity of interests is a major reason for the emergence and sustainability of residual-claimant business arrangements. But there is another reason why firms are commonly owned and managed by the same person, a reason that helps explain why the typical situation finds the boss hiring the workers instead of the workers hiring the boss.

People differ in a host of ways, and many of their differences have important implications for the type of productive efforts for which they are best suited. For example, both of the authors would have liked to have been successful movie stars, but because we have slightly less charisma than baking soda, we became economists instead. Had we been endowed with even less charm, we would have become accountants. More relevant to the current discussion, however, is the fact that people differ in their willingness to accept risk. Most people are what economists call *risk averse*; they shy away from activities whose outcomes are not known with reasonable certitude. Such people might, for example, prefer a sure $500 than a 50 percent chance of receiving $1,500 with a 50 percent chance of losing $500 (which has an expected value of $500).[34]

---

[33] The profits received by firms that are too large to be managed by single proprietors also serve to direct resources into their highest valued uses. But this is true because these firms are organized in ways that allow the owners (the residual claimants) to exert some control over those who manage the firm (the hired bosses). The problem that owners of large corporations face in controlling managers is discussed in subsequent chapters.

[34] The prevalence of insurance reflects the risk averseness of most people. Insurance allows people to experience a relatively small loss with 100 percent probability (their insurance premiums) in order to avoid a small chance of a much larger loss, but a loss with an expected value that is less than the insurance premiums. It is interesting to note, however, that the same people who buy fire insurance on their house will also buy lottery tickets. Buying a lottery ticket reflects risk-loving behavior since you are taking a small loss with 100 percent probability (the price of the lottery ticket) in order to take a chance on a payoff that is smaller in expected value than the loss. Explanations exist for why rational individuals would buy insurance and gamble. Probably the best known of these explanations was given by M. Friedman and L. J.

But some people are more risk averse than others, as measured by how much less than $500 a sure payoff would have to be before they would no longer prefer it to a gamble with a $500 expected value. And people who are highly risk averse will make very different career choices than those who are not.

Consider the choice between becoming a residual claimant by starting your own business and taking a job offered by a residual claimant. The choice to become a residual claimant is a risky one, requiring the purchase of productive capital and the hiring of workers (thereby obligating yourself to fixed payments) with no guarantee that the revenue generated will cover those costs. The person who starts a firm can lose a tremendous amount of money. Of course, in return for accepting this risk a residual claimant who combines keen foresight, hard work, and a certain amount of luck may end up claiming a lot of residual and becoming quite wealthy. Clearly, those willing to accept risks will tend to be attracted to a career of owning and managing businesses as residual claimants.

Those people who are more risk averse will tend to avoid the financial perils of entrepreneurship. They will find it more attractive to accept a job with a fixed and *relatively* secure wage, even though the return from such a job is less than the expected return from riskier entrepreneurial activity.

So business arrangements that put management control in the hands of residual claimants not only create strong incentives for efficient decisions, they also allow people to occupationally sort themselves out in accordance with an important difference in their productive attributes and their attitude toward risk. Not only will people who are not very risk averse be more comfortable as residual claimants than most people, they will generally be more competent at dealing with the risks that are inherent in organizing production in order to best respond to the constantly changing preferences of consumers. At the same time, those who are not averse to taking risks are likely less reliable at the relatively routine and predictable activity typically associated with earning a fixed wage than are those who are highly averse to risk.

By having people sort themselves into jobs according to their willingness to assume risk, the risk cost of doing business is minimized. And remember that when firms face competition in either their resource or product markets, they must look to lower all costs as much as possible. Otherwise, the firms' very existence can be threatened by those firms who pay attention to costs, including costs that are as hard to define as risk costs. If the firms that don't pay attention to costs avoid outright closure from being underpriced by competitors, they will be taken over by investors who detect an unexploited opportunity -- who buy the firms (or their stock) at a low price and sell them at a higher price after restructuring the firms to lower their costs.

Consider the prospect that more risk-averse workers own their firms and hire the less risk-averse owners of capital (as well as other resources) who would be paid a fixed

---

Savage ["The Utility Analysis of Choices Involving Risk," Journal of Political Economy , vol. 56 (August 1948), pp. 279-304]. But the fact remains than in situations that would put a significant amount of their wealth or income at risk, most people are risk averse.

return on their investments (with the fixed return having all the guarantees that are usually accorded worker wages).[35] Workers would then, in effect, be the residual claimants, and worker wages would then tend to vary (as do profits in the usual capitalist-owned firm) in less than predictable ways with the shifts in market forces and general economic conditions. Such a firm would not likely be a durable arrangement for even moderately large firms in which fixed investments are important. It's not hard to see why.[36]

The workers *might* be spurred to work harder and smarter because of the sense of ownership, which the proponents of worker ownership argue would be the case. But then, maybe not. Workers might be more inclined to shirk, given they are no longer pushed to work harder and smarter by owner-capitalists. And each worker can reason that his or her contribution to profits is very little (especially in large firms), so little that the power of residual claimacy is lost in the dispersion of ownership among workers. For this reason alone, we would expect most worker-owned firms to be relatively small.

Risk-averse worker-owners would require a "risk premium" built into their expected incomes, and their risk premium would be greater than the risk premium that the less averse owners of capital would require. Hence, the cost of doing business for the worker-owned firm would be higher than for the capitalist-owned firm, which means the worker-owned firms would tend to fail in competition with capitalist-owned firms. Instead of outright failure, we might expect many worker-owned firms to be converted to capitalist-owned firms simply because the workers would want to sell their ownership rights to the less risk-averse capitalists who, because of their lower risk aversion, can pay a higher price for ownership rights than other workers. The net income stream would be higher under the capitalist-owned firm, which means that the capital owners could pay more for the firm than it is worth to the workers. (The worker-owned firms would continue only if the workers were not allowed to sell their supposed ownership rights, which was true in the former Soviet Union and Yugoslavia.)

However, the worker-owned firm would be fraught with other competitive difficulties. Because of their risk aversion, workers would demand higher rates of return on their investments, a fact that would likely restrict their investments and lower their competitiveness and viability over the long run. Moreover, with workers are in control of the flow of payments to the capitalists after they, the capitalists, have made the fixed investment, the capitalists would have a serious worry. The capitalists must fear that the workers would tend to use their controlling position to appropriate the capital through non-competitive wages and fringe benefit payments to themselves, a fear that is not so prominent among workers when capitalists own the fixed assets and pay the workers a fixed wage.[37] Therefore, even the capitalists would require a risk premium before they invested in worker-owned firms.

---

[35] In effect, the owners of capital would hold financial assets that would have the look and feel of bonds.
[36] For an extended discussion of points in this section, see Michael C. Jensen and William H. Meckling, "Rights and Production Functions: An Application to Labor-Managed Firms and Codetermination," Journal of Business, vol. 52, no. 4 (1979), pp. 469-506.
[37] See the discussion of why workers do not own firms by Benjamin Klein, Robert G. Crawford, and Armen A. Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process,"

Of course, the workers could make the requisite investment, but we must wonder where they will obtain the investment funds. Out of their own pockets? Would they not want to put their own funds in secure investments? We must also wonder if workers would be interested in investing in their own worker-run firms. Like capitalists, workers can understand the threat to their investments from other workers, given the limited competitiveness of their worker-owned firms and the tendency of workers to restrict investment and drain the capital stock through over-payments in wages and fringes. Workers, however, have an additional problem: if they invest their financial resources in their own firms, then they will have a very narrow range of personal investments. By their work for their firms, they already plan to invest a great deal of their resources in their jobs just by spending time at work. Adding a financial investment means they will restrict the scope of assets in their personal portfolio of investments. That fact alone will increase their aversion to risky investments by their firm, and the longer the term of the investment, the greater the risk. Accordingly, we would expect the investments of worker-owned firms to be for shorter periods than would be the case in capitalist-owned firms, which implies that worker-owned firms would tend to lag in the development and application of new technologies. Such a tendency would once again make worker-owned firms less competitive, especially over the long run.

We are not suggesting that no firms will be worker-owned and managed. After all, some are. Instead, the analysis explains why there are relatively few such firms, and why they are typically small firms, relying primarily on human capital of the owner/workers rather than physical capital. When large firms, such as Weirton Steel and United Airlines, are worker-owned, they are not worker-managed. The worker-owners of such firms immediately hire bosses to make the tough decisions that have to be made to keep a firm viable, but then there are the inevitable tensions that come with worker ownership.

Worker-Owned Firms

Weirton Steel Company was taken over by employees in 1983. For a while it was a big success as workers put in long hours, helped each other outside their narrow work rule responsibilities, and did what it took so they could say "We kept the job moving," as maintenance worker Frank Slanchik said. But soon distrust built between workers and

---

Journal of Law and Economics, vol. 21 (1978), pp. 297-326. The problem of appropriation by workers is especially acute if the fixed assets are firm specific because they have no alternative use, which implies a limited resale value. As we have seen in other instances, owners of fixed assets with limited resale values open themselves to opportunistic behavior on the part of the buyer, in this case, the workers, who, once the specific investment is made, can appropriate the difference between the purchase and resale price. Workers hired by their capitalist-owners do not generally have the same worry about their work-related investments with their capitalist-owners. The workers' investments in their job-related skills are typically not firm specific. If workers need firm-specific skills, the workers can protect themselves from appropriation by having their firm pay for the investment they might make in firm-specific skills. Put another way, when human capital is relatively important on the job, we would expect the workers to also be the owners, which tends to be the case in accounting and law firms in which the ratio of human to physical capital investments tend to be high.

their managers (they still hire managers). The two big issues were money and management control. Slanchik notes, "These two issues are especially likely to crop up in capital-intensive industries such as steel and airlines, which constantly require huge capital expenditures that can be viewed as draining money away from potential wage increases."[38]

In July 1994, United Airline workers took an average pay cut of 15 percent for 55 percent interest in the company and 3 of its 12 seats on the airlines board of directors. According to Business Week, worker ownership of United Airlines has worked surprisingly well.[39] But even in the case of United, some problems that should have been expected are now evident. The 20,000 United flight attendants never joined the buyout and are still unhappy with the management. And, according to Business Week, "Many other employees still resent the pay cuts they took and suspect the ESOP [Employee Stock Ownership Plan] was foisted on them by greedy corporate executives and investment bankers who walked off with millions."[40] Moreover, the company offended many employees when it announced bonuses for 600 managers under a longstanding incentive-compensation plan. Investors have been reluctant to infuse additional capital into the airline, fearing that the employees would "revolt against cost-cutting decisions."[41]

This fear is so far unfounded, but the worker-ownership arrangement took place at the beginning of a very profitable period for airlines, United included. Part of the carrier's post-buyout success stems from a surge in air travel that has generated a record $2 billion in profits for the industry in 1996. Investors have to worry that when times get tougher in the future, United's newfound cooperative spirit might be seriously challenged, given that strains are already evident among the different worker groups. The 21,000 United Airlines flight attendants, who have been working without a contract for over a year, are thinking about an attack against United with a tactic known as "Create Havoc Around Our System" – or "Chaos."[42] The tactic consists of unannounced strike of individual flights, which can disrupt the entire schedule of an airline. Although the flight attendant union, the Association of Flight Attendants, says it does not want to invoke Chaos, but given United's "record profits," United attendants are "angry" and ready to strike, or so claims Kevin Lum, president of United's flight attendant association.[43]

Understandably, investors can't be sure just how tough United's workers will be on each other. They also have to fear that the workers would not add their share to the company's capital stock, by depleting retained earnings with wage increases, and would be tempted to drain the firm of any capital added by outside investors by way of wage

---

[38] Susan, Carey, "ESOP Fables: UAL Worker-Owners May Face Bumpy Ride If the Past Is a Guide," Wall Street Journal, December 23, 1993, p. 1.
[39] See Susan Chandler, "United We Own." Business Week March 18, 1996, pp. 96-100.
[40] Ibid., p. 98.
[41] Ibid., p. 99.
[42] In the WSJ on 24 June 1997 was an article by Susan Carey "United Flight Attendants Warn of 'Chaos'," Wall Street Journal, pp. B-1 and B-2.
[43] Ibid.

increases. The workers have to worry about the inclination of each worker group to garner firm profits at the expense of other groups and the investors. The workers also have to worry that they have taken over the role of the investors, which is accepting the risk that comes from being residual claimants. The workers' insecurities can be heightened by the fact that the company's future will be jeopardized by the absence of the capital that it will need to remain competitive with investor-owned airlines that don't have the problems and fears that United might have.

We should not be surprised if, at some later date, the workers effectively try to "buy back" some security by selling their stake in their company, giving the investors that right to be tough bosses in exchange for more investment funds and a more certain income stream for workers (with more of their income coming from wages, salaries, and fringe benefits and less of it coming from dividends).

Management Snooping

Technology has given workers a chance to loaf on the job while they appear busy at their desk. All workers have to do is surf the web for entertainment, shopping, and sex sites on their office computers while giving passersby (including their bosses) the impression that they, the workers, couldn't be more focused on company business. And workers are often good at acting busy and engaged.

At the same time, technology is coming to the rescue of manager/monitors – or bosses who want to be really tough, if not oppressive. Programs such NetNanny, SurfWatch, and CyberPatrol enable managers to block worker access to web sites with certain words on the site, for example, "sex." However, with the aid of a program called com.Policy from SilverStone Software, managers now can, from their own desktop computers, go much further and check out what worker's have on their computer screens. The software can take a snapshot of the worker's computer screen and sends it, via the local area network, to the boss' screen. If a worker visits an XXX-rated web site or writes a love note to a coworker or someone across the country, managers can know it, and, depending on how tough they want to be, the managers can penalize or dismiss the workers for using company equipment for personal use. Presumably, the managers can, with the aid of the software, increase worker productivity, given that the penalties or threat of penalties, can eliminate worker shirking.

The real question is Should managers use technology that allows them to "snoop" (to use the characterization of the technology's critics)? Would workers want them to use it? Clearly, there are good reasons managers and workers alike would not want to use the software, it represents an invasion of worker privacy. Many managers and, we suppose, almost all workers, find "snooping" distasteful. But, as in all other business matters, the worker problems must be weighed off against the benefits to the firm *and* workers.

Workers might not want their privacy invaded at the whim of their bosses, but the workers can understand the now familiar prisoner's dilemma they are in -- one in which many of the workers might be inclined to misuse their office computers for private gain (entertainment, maintenance of love affairs, and sexual stimulation). In large offices, the

workers can reason that everyone else is misusing (at least to some extent) their computers, that their individual misuse will have an inconsequential impact on the firm's profitability or survivability, and that they each worker should do what everyone else is doing, take advantage of the opportunity to misuse their computers – even though long-run firm profits and worker wages will suffer as a result of what the workers do (or, rather, don't do).

Accordingly, workers could welcome the invasion of their privacy, primarily because the gain in income and long-term job security is of greater value than the loss of privacy. Managers can use the software simply because they are doing what their stockholders *and* workers want them to do, make mutually beneficial trades with their workers, which is, in this case, ask them to give up some privacy in exchange for the prospects of higher wages and security.

At the same time, we should not expect that the above deduction will apply in every worker group. Some worker groups will value their privacy very highly, so highly in fact that in some instances the managers would have to add more to worker wages than the firm could gain in greater productivity from use of the monitoring software. In such cases, use of the software would be nonsensical: it would hurt both the workers and the firm's bottom line. Put another way, some bosses aren't as tough as they might want to be simply because, beyond some point, toughness – added "snooping" -- doesn't pay; it can be a net drain on the company.

Critics of the snooping software are prone to characterize it as "intrusive," if not "Orwellian." One such critic was reported to have reacted to the software's introduction with the comment, "It worries me that with the assistance of a variety of tools that every moment of a person's workday can be monitored. Workers are not robots that work 24 hours a day without ceasing."[44] We simply don't see the matter in such black and white terms. The old quip "different strokes for different folks" contains much wisdom, especially in business. We see nothing wrong with employers warning their employees, "The computers are the firm's, and we reserve the right to snoop on what you are doing with the firm's equipment as we see fit." To the extent that the (potential) snooping is seen as a threat to workers, the firm would have to pay in higher wages for the snooping bosses might do. If they did not pay a higher wage for the announcement, workers could be expected to go elsewhere, where the firm explicitly rules out snooping. What is understandably objectionable to employees is the snooping when it is not announced or, worse yet, when managers profess, or just intimate, that they will not use the available technology, but then snoop at will. Such managers not only violate the privacy and trust of their workers, they engage in a form of fraud. They effectively ask their workers to take a lower rate of pay than they would otherwise demand, and then don't give their workers what they pay for, privacy. Moreover, such after-the-fact snooping doesn't do what the firm wants, increase *beforehand* the incentive workers have to apply themselves.

---

[44] As quoted in Lisa Wirthman, "Superior Snooping: New Software Can Catch Workers Goofing Off, But Some Say Such Surveillance Goes Too Far," Orange County (Calif.) Register, July 20, 1997, p. 1 and 10 (connect section).

Unannounced snooping is just poor management policy on virtually all scores. With announced snooping policies, workers can sort themselves among firms. Those workers who value their privacy or on-the-job entertainment highly can work for firms that don't snoop. Those workers who value their privacy very little can work for firms that announce that they might snoop. "Different strokes for different folks" can be a means of elevating on-the-job satisfaction.

What firms would be most likely to use the monitoring software (or any other technology that permits close scrutiny of worker behavior)? We can't give a totally satisfactory answer. Workplace conditions and worker preferences are bound to vary across industries. But we can say with conviction that there is no "one-size/fits-all" monitoring policy. We can only imagine that different firms will announce different levels of snooping -- with some firms ruling it out, other firms adopting close snooping, and still others announcing occasional snooping. And many firms with the same level of snooping can be expected to impose penalties with different levels of severity.

Although we can't say much in theory about what firms should do, we can note that the snooping software, and similar technologies, would more likely be used in "large" firms where the output of individual workers is hard to detect, measure, and monitor than in "small" firms where output is relatively easy to detect, measure, and monitor precisely because each worker's contribution to firm output is such a large share of the total. The snooping technology would not likely be used among workers whose incomes are tied strongly to measures of their performance, for example, sales people who are on commission and far removed from the company headquarters. Such workers will suffer a personal cost if they spend their work time surfing the web or writing love notes. Managers should be little more concerned with such workers' misuse of their company computers than they are concerned about how their workers use their paychecks at the mall. If such workers are not performing (because they are "spending" too much of their pay on net surfing), then the firm should consider the prospect that they need to increase the cost of wasted time by more strongly tying pay to performance (a subject to which we return in a later chapter).

By implication, managers will not likely use the software to monitor employees who are highly creative. "Creativity" does not always happen when workers diligently apply themselves, and often occurs precisely because workers are relaxed, with the ability to do as they pleased without fear of being penalized for goofing off. Firms would probably be more inclined to use the software with employees who are paid by the hour and have little or no personal payoff from working hard and smart. It should go without saying that the more workers value their privacy, the less likely monitoring software will be used. This is because the more workers value their privacy, the more managers would have to pay in higher wages to invade the privacy.

The Reason for Corporations

Competition determines which business arrangements will survive and which will not. The prevalence of single proprietorships is explained by the advantage of this business

form in producing those products the consumers want as inexpensively as possible. But changing circumstances can reduce the competitive advantage of a business arrangement as new arrangements are found to do a better job of organizing productive activity. Technological advances that took place during the latter part of the nineteenth century made it possible to realize huge economies from large scale production in many manufacturing industries. These technological advances shifted the advantage to business organizations that were far too large to be owned and managed by one proprietor, or even by a few. But the advantage of large business firms is reduced by the fact that they make it impossible to concentrate the motivation created by ownership entirely in the hands of those making management decisions.

Those manufacturing firms that developed organizational arrangements that did the best job of reducing the disconnection between the owners' incentives and the managers' control were best able to take advantage of economies from large-scale production. The result was a competition that resulted in the development of the modem corporation, the business form that today accounts for most of the value produced in the United States economy, even though small owner-managed firms still make up, by far, the largest number of firms in the economy.

However, it must be remembered (contrary to what is often taught in business books) that the corporation (an organization under which investors have limited liability) was not a creation of the state.[45] The corporation emerged before states got into the incorporating business. Groups of private investors formed corporations because they believed that there were economies to be had if they all agreed to create a business in which outside parties could not hold the individual investors liable for more than their investment in the corporation (that is, the investors' personal fortunes would not be at risk from the operation of the firm, as was and remains true of proprietorships and partnerships). Clearly, such a public announcement of limited liability (made evident with "Inc." on the end of corporate names) might make lenders weary and cause them to demand higher interest rates on loans. However, the firm would have the offsetting advantage of being able to attract more funds from more investors, increasing firm equity, a force that could not only increase the firm's ability to achieve scale economies grounded in technology, but would lower risk costs to lenders. Of course, the outside investors could be hard taskmasters, given that they could shift their investment away from firms not maximizing profitably. But that doesn't mean the workers would find the corporate form unattractive. On the contrary, given the potential scale economies and risk reductions, corporations may provide more secure employment than small proprietorships.

Jack Welch, the chief executive officer of General Electric, has played out the central point of this "Manager's Corner" because he surely qualifies as a tough boss.

---

[45] Robert Hessen, In Defense of the Corporation (Stanford, Calif.: Hoover Institution Press, 1979) develops this view of the corporation.

Indeed, Fortune once named Welch "America's Toughest Boss."[46] Welch earned his reputation by cutting payrolls, closing plants and demanding more from those that remained open. Needless to say, these decisions were not always popular with workers at GE. But today, GE is one of America's most profitable companies, creating far more wealth to the economy and opportunities for its workers than it would have if the tough and unpopular decisions had not been made. In Welch's words, "Now people come to work with a different agenda: They want to win against the competition, because they know that . . . . customers are their only source of job security. They don't like weak managers, because they know that the weak managers of the 1970s and 1980s cost millions of people their jobs."[47]

MANAGER'S CORNER II: **The Value of Teams**

The central reason firms exist is that people are often more productive when they work together -- in "teams" -- than when they work in isolation from one another but are tied together by markets. "Teams" are no passing and empty management fad. Firms have always utilized them. What seems to be new is the emphasis within management circles on the economies that can be garnered from assigning complex sets of tasks to relatively small teams of workers, those within departments and, for larger projects, across departments. However, "teams" also present problems in the form of opportunities for shirking (which should be self–evident to many MBA students who form their own study and project groups to complete class assignments). A central problem managers face is constructing teams so that they minimize the amount of shirking.

At its defense avionics plant, Honeywell reports that its on-time delivery went from 40 percent in the late 1980s to 99 percent at the start of 1996, when it substituted teams, in which workers' contributions are regulated by the members, for assembly-line production, in which workers' contributions are regulated extensively by the speed of the motors that drive the conveyor belts. Dell Computer is convinced that its team-based production has improved quality in its made-to-order mail-order sales. Within twelve months of switching to teams in its battery production, a different company, Electrosource, found its output per worker doubled (with its workforce dropping from 300 to 80 workers).[48]

If people could not increase their joint productivity by cooperating, we would observe individual proprietorships (with no employees other than the owners) being the most common form of business organization and also the form that contributed most to national production. As it is, while proprietorships outnumber other business forms (for example, partnerships and corporations) by a wide margin, they account for only a minor fraction of the nation's output. Even then, many proprietorships can't get along without a few employees. Single-worker firms tend to be associated with the arts. Few artists have

---

[46] Noel M. Tichy and Stratford Sherman, "Jack Welch's Lessons for Success," Fortune (25 January 1993) pp. 86-93.
[47] Ibid. p. 92.

[48] As reported in Paulette Thomas, "Work Week: Teams Rule," Wall Street Journal, May 28, 1996, p. A1.

employees. Even we are writing this book as a partnership in the expectation that our joint efforts will pay off in a better book than either of us could write alone. We are a "team" of a sort. But notice there are only two of us, and we aren't about to write a book with a number of others, for reasons explained below. As important as teams can be in business, managers must recognize inherent incentive problems that limit the size of productive teams.

Team Production

To be exact, what do we mean by "team production"? If Mary and Jim could each produce 100 widgets independent of one another and could together produce only 200 widgets, there would be no basis for team production, and no basis for the two to form a firm with all of the trappings of a hierarchy. The added cost of their organization would, no doubt, make them uncompetitive *vis a vis* other producers like themselves who worked independently of one another. However, if Mary and Jim could produce 250 widgets when working together, then team production might be profitable (depending on the exact costs associated with operating their two-person organization).

Hence, we would define **team production** as those forms of work in which results are highly interactive: The output of any one member of the group is dependent on what the other group members do. The simplest and clearest form of "team work" is that which occurs when Mary and Jim (and any number of other people) move objects that neither can handle alone from one place to another. The work of people on an assembly line or on a television-advertising project is a more complicated form of teamwork.

Granted, finding business endeavors that have the potential of expanding output by more than the growth in the number of employees is a major problem businesses face, but it is not the only problem and may not be the more pressing day-to-day problem when groups of people are required to do the work. The truly pressing problem facing managers on a daily basis is making sure that the synergetic potential of the workers who are brought together into a team is actually realized, that is, production is carried out in a cost-effective manner, so that the cost of organization does not dissipate the expanded output of, in our simple Mary/Jim example, 50 widgets.[49]

We often think of firms failing for purely financial reasons. They don't make a profit, or they incur losses. Firms are said to be illiquid and insolvent when they fail. That view of failure is instructive, but the matter can also be seen in a different light, as an organizational problem *and* a failure in organizational incentives. A poorly run organization can mean that all of the 50 "extra" widgets that Mary and Jim can produce together are lost in unnecessary expenditures and impaired productivity. If the organizational costs exceed the equivalent of 50 widgets, then we can say that Mary and Jim have incurred a loss, which would force them to adjust their practices as a firm or to part ways.

[49]We remind the reader that "cost" is the value of that which is foregone when something is done. Cost can be measured in money, but the real cost is the value of that which is actually given up.

Many firms do fail and break apart, not because the *potential* for expanded output does not exist, but because the potential is not realized when it could be. The people who are organized in the firm can do better apart, or in other organizations, than they can together. That's what we really mean by *reoccurring* business "losses."

Why can't people always realize their collective potential? There is a multitude of answers that question. Firms may not have the requisite product design or a well-thought-out business strategy to promote the products. Some people just can't get along; they rub each other wrong when they try to cooperate. Nasty conflicts, which deflect people's energies at work to interpersonal defensive and predatory actions, can be so frequent that the production potentials are missed.

While recognizing many non-economic explanations for organizational problems, we, however, would like to stay with our recurring theme, that incentives always matter a great deal and they can become problematic within firms. Our general answer to our question, why firms' potential can go unrealized, is that frequently the firm does not find ways to properly align the interests of the workers with the interests of other workers and the owners. They don't cooperate like they should.

In our simple firm example, involving only two people, Mary and Jim, each party has a strong *personal* incentive (quite apart from an altruistic motivation) to work with the other. After all, Mary's contribution to firm output is easily detected by her *and* by Jim. The same is true for Jim. Moreover, each can readily tell when the other person is not contributing what is expected (or agreed upon). Each might like to sit on his or her hands and let the other person carry the full workload. However, the potential is not then likely to be realized, given that the active participation of both Mary and Jim is what generates the added production and their reason for wanting to become a firm (or team) in the first place.

Furthermore, Jim can tell when Mary is shirking her duties, and vice versa, just by looking at the output figures and knowing that there is only one other person to blame. Accordingly, when Mary shirks, Jim can "punish" Mary by shirking also, and vice versa, ensuring that they both will be worse off than they would have been had they never sought to cooperate. The agreement Mary and Jim might have to work together can be, in this way and to this extent, self-enforcing, with each checking the other -- and each effectively threatening the other with reprisal in kind. The threat of added cost is especially powerful when Mary and Jim are also the owners of the firm. The cost of the shirking and any "tit for tat" consequences are fully borne by the two of them. There is no prospect for cost shifting.

Two-person firms are, conceptually, the easiest business ventures to organize and manage because the incentives are so obvious and strong and properly aligned. Organizational and management problems can begin to mount, however, as the number of people in the firm or "team" begins to mount.

Everyone who joins a firm may have the same objective as Mary and Jim -- they all may want to make as much money as possible, or reap the full synergetic potential of their cooperative efforts. At the same time, a number of things can happen as the size of

the firm or "team" grows in terms of more employees. Clearly, communication becomes more and more problematic. What the boss says can become muffled and less clear and forceful as the message is spread through more and more people within the firm.

Also, and probably more importantly, as explained in the "logic of group behavior," incentives begin to change with the growth in the size of groups. Foremost, each individual's contribution to the totality of firm output becomes less and less obvious as the number of people grows. This is especially true when the firm is organized to take advantage of people's specialties. Employees often don't know what their colleagues do and, therefore, are not able to assess their work.

When Mary is one of two people in a firm, then she is responsible for half of the output (assuming equal contributions, of course), but when she is one of a thousand people, her contribution is down to one-tenth of one percent of firm output. If she is a clerk in the advertising department assigned to mailing checks for ads, she might not even be able to tell that she is responsible for one-tenth of a percent of output, income, and profits.

If Mary works for a firm with several hundred thousand workers, you can bet that she has a hard time identifying just how much she contributes to the firm. She can't tell that she is contributing anything at all, and neither can anyone else. She can literally get lost in the company. If she doesn't contribute, she and others will have an equally difficult time figuring out what exactly was lost to the firm. Her firm's survival is not likely to be materially affected by what she does or does not do. She is the proverbial "drop in the bucket," and the bigger the bucket, the less consequential each drop is. Of course, the same could be said of Jim and everyone else in the firm.

Now, it might be said that all of the "drops" add up to a "bucket." The problem is that each person must look at what he or she can do, given what all the others do. And drops, taken individually, don't really matter, so long as there are a lot of other drops around.

Admittedly, if no one else contributes anything to production (there are no other drops in the bucket), the contribution of any one person is material -- in fact, everything. The point is that in large groups and as output expands, each worker has an *impaired* incentive to do that which is in all of their interests to do -- that is, to make their small contribution to the sum total of what the firm does. All workers may want the bucket to get filled, but to do so takes more than wishful thinking, which often comes in the form of assuming that people will dutifully do that which they were hired to do. The point here is that large-number prisoners' dilemmas are more troublesome than small-number prisoners' dilemmas.

A central lesson of this discussion is, as stressed before, not that managers can never expect workers to cooperate. We concede that most people do have – very likely because of genetics and the way they were reared -- a "moral sense," or capacity to do what they have committed to doing -- that they will cooperate, but only *to a degree, given normal circumstances*. However, there are countervailing incentive forces embedded in the way groups – or teams – of people work that, unless attention is given to the details of

firm organization, can undercut the power of people's natural tendencies to cooperate and achieve their synergetic potential. If people were total angels, always inclined to do as they are told or as they said they would do, then the role of managers would be seriously contracted. Even if almost everyone were inclined to do as they were told or committed to doing, still managers would want to have in place policies and an organizational structure that would prevent the few "bad" people from doing real damage to the firm, which, if left unchecked, they certainly could do. The arguments presented also help us answer several questions.

Why are there so many small firms? Many commentators give answers based on technology: Economies of scale (relating strictly to production techniques and equipment) are highly limited in many industries. One very good organizational reason is that many firms have not been able to overcome the disincentives of size, making expansion too costly and uncompetitive.

Why are large firms broken into departments? While it might be thought that the administrative overhead of department structures, which requires that each department have a manager and an office with all the trappings of departmental power, is "unnecessary," departments are a means firms use to reduce the size of the relevant group within the firm. The purpose is not only to make sure that the actions of individuals can be monitored more closely by bosses, but also that the individuals in any given department can more easily recognize their own and others' contributions to "output."

Why do workers have departmental bosses? One reason is that the owners want their instructions to be carried out. Another explanation, one favored by UCLA economists Armen Alchian and Harold Demsetz, is that the workers themselves want someone who is capable of monitoring the output of their co-workers, to prevent than from shirking and to increase the incomes and job security of all workers.[50] Workers want someone who is given the authority to fire members who shirk. As discussed under "Manager's Corner I," if owners didn't create bosses, then the workers probably would want them created in many situations for many of the same reasons and from much the same mold as do owners.

Why is there so much current interest in "teams"? As acknowledged, we suspect that the concept of teams in industry has always been around and used for a long time. After all, we have worked as members of "teams" (mainly, departments of business and economics professors) for all of our careers. However, it is also likely that over recent decades, managers probably became far too enamored with the dictates of "scientific management," which focused on the means of controlling workers with punishments and rewards that come from bosses who are outside (and above) the workers' immediate working group. Managers tried with some success to reduce shirking with the introduction of the assembly lines, under which the speed of the assembly-line belt determined how fast workers worked (with the presumption that workers would not have much leeway to adjust behavior, which might have been true for the pace of the work

[50] Armen Alchian and Harold Demsetz, "Production, Information Costs, and Economic Organization," American Economic Review, vol. 62 (1972), pp. 777-795.

done but not the quality). In the past, many managers have overlooked the impact of team size on member incentives. They have now begun to realize that they can increase worker productivity by reducing the size of the relevant group, to ensure that workers, who know most about what needs to be done in many firms, can monitor each other. Workers in appropriately sized teams can monitor and direct each other's work. Such close-at-hand monitoring can become even more important when consumers begin to demand more emphasis on quality, as they already have.

We also suspect that the modern interest in "teams" is driven by newfound global competition and by the growing sophistication of work in many industries. Those firms, domestic and foreign, that have employed teams successfully have forced other firms with traditional top-down management/control structures to also consider teams to keep up with the competition. Technology has greatly elevated the sophistication of production, increasing the specialization of work with much of the knowledge of what can be done in production known only by the people who actually do the jobs. Bosses can know a lot, but they can't possibly know many of the things that their workers know. Managers must delegate decision-making authority to those who have the detailed knowledge to make the most cost-effective decisions, which, when production is interdependent or done jointly by a number of people, means decisions must be made by teams of workers.

As a consequence of the benefits of team production, we should not be surprised that at Motorola's Arlington Heights cellular phone plant, team members participate in the hiring and firing of co-workers, determine training, and set work schedules. At Nucor Steel, teams can discipline their members. At both companies, the team-based plants are remarkably productive.

At the same time the team members are delegated decision-making authority, they must also shoulder responsibility for the decisions they make. That necessarily means that team members must share the rewards from good decisions and the costs from bad ones. Often, this can mean that production bonuses are tied to what the team *as a whole* accomplishes, not what individual members do. Often, it also means that when the decisions are systematically bad, then the entire team must be dismissed, not just individuals. If individuals can be chosen as scapegoats for the actions taken by their team, then all individuals will have an incentive to "game" the process, trying to shirk and then pinning the blame on others. Team members will then have less incentive to work together and more incentive for political intrigue, possibly corrupting the working relationship of all.

A natural question that is bound to puzzle business managers interested in maximizing firm output is, How large should teams be? How many members should they have? We obviously can't say exactly, given the many factors that explain the great variety of firms in the country. (If we could formulate a pat answer, this book would surely sell zillions!) However, we can make several general observations, the most important of which is that managers must acknowledge that shirking (or "social loafing") will tend to rise along with the size of the group, everything else held constant.

In addition, we suggest that since people who are alike tend to cooperate, the more alike the members, the larger the team can be. The more training team members are given in cooperation, the larger the teams can be. Training, in other words, can pay not only because it makes workers more productive, given how much the workers know how to do, but also because it can reduce the added overhead of a larger number of smaller departments.

However, a lot depends on the type of training given workers. Apparently, economists, using their maximizing models (and the firmly held belief that everyone will shirk when they can), are inclined to play whatever margins are available to their own personal advantage, or to shirk when feasible, to a degree not true of other professionals.[51] As a consequence, it probably follows that the more economists (and other people with similar conceptual leanings) employed, the smaller the team can be. Although we may never have intended it, we must fear that the people who read this book may be less disposed to cooperate than they were before they picked it up.

The more workers are imbued with a corporate culture and accept the firm's goals, the larger the team can be. The expenditures by corporate leaders trying to define the firm's purpose can be self-financing, given that the resulting larger departments can release financial and real resources.

The more detectable or measurable are the outputs of individual team members by other team members, the larger the team can be. Firms, thereby, have an economic interest in developing ways to make work, or what is produced, objective. Finally, the greater the importance of quality, the more important team production should be, and the smaller teams will tend to be.

No matter how it is done, the size of the teams within a firm can affect the overall size of the firm. Firms with teams that are "too large" or "too small" can have unnecessarily high cost structures that can restrict the firms' market shares and overall size, as well as the incomes of the workers and owners.

But recognizing that teams can add to firm output is only half the struggle to achieve greater output by getting workers to perform as they should. A question that all too often undercuts the value of teams is, "How are the workers to be paid?" If workers are rewarded only for the output of the team, then individual workers once again have incentives to "free ride" on the work of others (to the extent that they can get away with it, given the size of the team), which can be realized in not only slack work, but also absenteeism. If team members are rewarded exclusively for their own individual contributions, then the incentive is reduced for actual teamwork.

---

[51] Researchers have found that on single-play experimental games designed to test the tendency of people to "free ride" on the group's efforts, not everyone contributed to the group's output. However, they also found that the members produced 40 to 60 percent of the "optimal output" of the public good, with the exception of only one notable group, graduate students in economics. These graduate students provided only 20 percent of the optimal output. See Gerald Marwell and Ruth Ames, "Economists Free Ride, Does Anyone Else?" Journal of Public Economics, vol. 15 (1981), pp. 295-310.

Generally managers effectively "punt" on compensation issues, not knowing exactly how to structure rewards, by offering compensation that is based partly on team output and partly on individual contributions to the team. Team output is generally the easier of the two compensation variables to measure, given that the teams are organized along functional lines, with some measurable objective in mind. Individual contributions are often determined partially by peer evaluation, given that team members are the ones who have localized knowledge of who is contributing how much to team output. But here again, the compensation problem is not completely solved. Team members can reason that how they work and how they and their cohorts are evaluated can affect their slice of the compensation pie. The greater the evaluation of others, the lower their own evaluation, a consideration that can lead team members to underrate the work of other team members. The result can be team discord, as has been the experience at jean maker Levi-Strauss where supervisors reportedly spend a nontrivial amount of time refereeing team-member conflicts. To ameliorate (but not totally quell) the discord, Levi-Strauss has resorted to giving employees training in group dynamics and methods of getting along.[52]

Motivating Team Members

One of the questions our conceptual discussion cannot answer totally satisfactorily is, "How can managers best motivate workers to contribute to team output?" There are four identifiable pay methods worth considering:

1. The workers can simply share in the revenues generated by the team (or firm). We can call this reward system *revenue sharing*. The gain to each worker is the added revenue received minus the cost to the worker of the added effort expended. Under this method reward, each worker has maximum incentive to free ride, especially when the "team" is large.
2. The workers can be assigned target production or revenue levels and be given what are called *forcing contracts*, or a guarantee of one high wage level (significantly above their market wage) if the target is achieved and another, lower (penalty) wage if the target is not achieved. Under this system, each worker suffers a personal income loss from the failure of the team to work effectively to meet the target.
3. The workers can also be given an opportunity to share in the team or firm profits. Profit sharing (or sometimes called "gainsharing") is, basically, another form of a forcing contract, since the worker will get one income if the firm makes a profit (above some target level) and a lower income if the profit (above a target level) is zero.
4. The workers within different teams can also be rewarded according to how well they do relative to other teams. They can be asked to participate in *tournaments*, in which the members of the "wining team" are given higher

[52] As reported in R. Mitchell, "Managing by Values," Business Week, August 1, 1994, p. 50.

> incomes -- and, very likely, higher rates of pay by the hour or month -- than the members of other teams. We say "very likely" because the winning team members may work harder, longer, and smarter in order to win the tournament "prize." Hence, the "winner'" pay per hour (or any other unit of time) could be lower than the "losers."[53]

All of the pay systems may have a positive impact on worker input and, as a consequence, on worker output. For example, a number of studies reveal that profit sharing and worker stock ownership plans do seem to have an impact on worker productivity.[54] One study of 52 firms in the engineering industry in the United Kingdom (40 percent of which had some form of profit-sharing plans and the rest did not) found that profit sharing could add between 3 and 8 percent to firm productivity.[55] And it has also been shown that the more "participatory" the decision-making process, the more the information-sharing the communication process, the more flexible the job assignment, and the greater the extent of profit sharing, the greater worker performance relative to more traditional organizational structures.[56] But the question that has all too infrequently been addressed is which method of rewarding workers and their teams is *more* effective in overcoming shirking and causing workers to apply themselves?

One of the more interesting studies that addresses that question uses an experimental/laboratory approach to develop a tentative assessment of the absolute and relative value of the different pay methods on worker effort. Experimental economists Haig Nalbantian and Andrew Schotter used two groups of six university economics students in a highly stylized experiment in which the students' pay for their participation

---

[53] We should not be surprised if the *pay rates* of the winning and losing teams are closer together than their *incomes*. We doubt, however, a pay system that resulted in the "winners" having a lower rate of pay than the "losers" would for long have the desired incentive impact, given that the higher income must also be discounted by the probability of any team winning." If the winners' pay rate were not higher than the losers', we would expect the winners to curb their effort.

[54] See Felix FitzRoy and Kornelius Kraft, "Profitability and Profit-Sharing," Journal of Industrial Economics, vol. 35 (no. 2) December 1986, pp. 113-130; Bion B. Howard and Peter O. Dietz, A Study of the Financial Significance of Profit Sharing (Chicago: Council of Profit Sharing Industries, 1969); Bertram L. Metzger, Profit Sharing in 38 Large Companies, I & II (Evanston, Ill.: Profit Sharing Research Foundation, 1975); Bertram L. Metzger and Jerome A. Colletti, Does Profit Sharing Pay? (Evanston, Ill.: Profit Sharing Research Foundation, 1975); John L. Wagner, Paul A Rubin, and Thomas J. Callahan, "Incentive Payment and Non-Managerial Productivity: An Interrupted Time Series Analysis of Magnitude and Trend," Organizational Behavior and Human Decision Processes, vol. 42 (no. 1), August 1988, pp. 47-74; Martin L. Weisman and Douglas L. Kruse, "Profit Sharing and Productivity," in Alan S. Blinder, ed., Paying for Productivity: A Look at the Evidence (Washington, D.C.: Brookings Institution, 1990), pp. 95-140: and U.S. Department of Labor, High Performance Work Practices and Firm Performance (Washington, D.C.: U.S. Government Printing Office, 1993).

[55] John Cable and Nicolas Wilson, "Profit-Sharing and Productivity: An Analysis of UK Engineering Firms," Economic Journal, vol. 99 (June 1989), pp. 366-375.

[56] See Mark Husled, "The Impact of Human Resource Management Practices on Turnover, Productivity and Corporate Financial Performance," Academy of Management Journal, vol. 38 (no. 2), June 1995, pp. 635-672; and Casey Ichniowski, Kathryn Shaw, and Giovanna Prennushi, The Effects of Human Resource Practices on Productivity (Cambridge, Mass.: National Bureau of Economic Research, working paper no. 5333, 1996).

in the experiment would be determined by how "profitable" their respective teams were in achieving maximum "output."[57]

The students did their "work" on computers that were isolated from one another. The students indicated how much "work" they would do in the 25 rounds of the experiment by selecting a number from 0 to 100 that had a cost tied to it, and each higher number had a higher cost to the student, just as rising effort tends to impose an escalating cost on workers. The students in each of the two teams always knew two pieces of important information, how much they "worked" (or the number they submitted) in each round and how much the "team" as a total "worked." They did not know the individual "effort levels" of the other students.

Granted, there is much to be desired about the experiment, which the authors fully conceded. The experimental setting did not reflect the full complexity of the typical workplace. Direct communication among workers can have an important impact on the effort levels of individual workers, but the complexity of the workplace is why it is so difficult to determine how pay systems affect worker performance, especially relative to alternative compensation schemes.

Nonetheless, the researchers were able to draw conclusions that generally confirm expectations from the theory at the heart of this book. They found that when the revenue-sharing method of pay was employed, the median "effort level" for each of the two teams started at a mere 30 (with a maximum effort level of 100), but since the students were then told how little effort other team members were expending in total, the students began to cut their own effort in each of the successive rounds. The median effort level in both teams trended downward until the 25th round when the median effort level was under 13. That finding caused the researchers to assert: "Shirking happens."[58] They were also able to deduce that the history of the team performance matters: the higher the team performance at the start, the greater the team performance thereafter (although the effort level might be declining over the rounds, it would still be higher at identified rounds, the higher the starting effort level).

Nalbantian and Schotter found that forcing contracts and profit sharing could increase the initial level of effort to 40 or above, a third higher than the initial effort level under revenue sharing, but still the effort level under forcing contracts and profit sharing trended downward with succeeding rounds of the experiment. Nalbantian and Schotter also found that the tournaments that were tried, which forced the team members to think competitively, had median initial effort levels on par with the initial effort levels observed under forcing contracts. However, the effort level tended to increase in the first few round and then held more or less constant through the rest of the 25 rounds. At the end of the 25 rounds, the teams had a median effort level of 40 to 50, or up to four times the ending effort level under the revenue-sharing incentive system. Understandably, the authors conclude that "a little competition goes a very long, long way."[59]

---

[57] Haig R. Nalbantian and Andrew Schotter, "Productivity Under Group Incentives: An Experimental Study," American Economic Review, vol. 87 (no. 3), June 1997, pp. 314-341.
[58] Ibid., p. 315.
[59] Ibid.

Finally, the authors conclude that monitoring works, which is no surprise, but the extent to which monitoring hiked the effort level does grab the attention. No monitoring system works perfectly, so the authors evaluated how the teams would perform with a competitive team pay system under two experimental conditions, one in which the probability of team members being caught shirking was 70 percent of the time and one in which teams members being caught shirking was 30 percent of the time, with the penalty being stiff, loss of their "jobs." The median effort for one team level started about at 75 (the predicted effort level from theory) and stayed there until the last round, at which point the effort level fell markedly (a finding that will be understandable from our discussion of the "last-period problem" in an earlier chapter). The median effort level for the other team started at about 50, rose quickly to 70, and stayed there through the rest of the rounds (with one very large drop in effort in the middle of the rounds).

When the probability of being caught shirking dropped to 30 percent, the effort level of one team started at 70 and went up and down wildly between zero and 80 for the next twenty rounds, only to approach zero during the last five rounds. The effort level of the other team started close to zero and stayed very close to zero for most of the following rounds (reaching above 10 only twice).

Obviously, monitoring of team members can have a dramatic impact on team performance, but as in all matters, the cost of the monitoring system can be high. The researchers have not yet been able to say, from the experimental evidence, whether the improvement in team performance is worth the cost of the monitoring system that is required. However, managers can't wait for the experimental findings. They must find ways of minimizing the monitoring costs. One of the great cost-saving advantages of teams, which is not reflected in the way the experiments were run, is that teamwork tends to be self-monitoring, with each team member monitoring one other. In the experiment, the team members could not monitor and penalize each other. When the experimental work is extended, we would not be surprised if the effort level increased when the team members are able to monitor and penalize each other.

Should all firms adopt the competitive team approach? The evidence suggests a firm "yes." But we hasten to add a caveat that managers of some firms must keep in mind. Greater effort to produce more output is desirable so long as it does not come with a sacrifice in "quality" (or some other important dimension of production). Competitive team production may be shunned in firms in industries like pharmaceuticals and banking that can't tolerate, because, for example, of liability problems, concessions in their quality standards. The competition in the tournaments drive up "output" but drive down "quality." Such firms would want to use reward systems that keep the competition under control and the quality standards up. They would also want to rely on close monitoring, and they could justify the cost, given the costs that they might suffer with defects. This leads to the obvious conclusion, the greater the cost of mistakes, the greater the cost that can be endured from relaxed competition and from monitoring.

Problems with Committees

Committees are special forms of teams that are the subject of much business abuse, both in terms of the number of meetings held and in terms of what business people think of most meetings. Indeed, business people often chafe when the subject of committee meetings is aired, "People talk too much, and too little is accomplished," harried businessmen and women often fret about committee meetings. By the standards of university faculty meetings, however, business people have nothing to complain about. Indeed, they can thank their lucky stars that they do not have to suffer many of the meetings we've had to suffer throughout our academic careers. Now, we have something to complain about! Business people may talk a lot, but faculty members have made "hot air" an entitlement.

Why are committee meetings so boring, as well as frequently unproductive? We suspect that the problem emerges partly because the people who call the meetings do not necessarily suffer the costs that are incurred. We were once listening to a business executive give a talk in which he crystallized his point, and ours, by asking the audience for a show of hands in response to the question, "How many people in this room can sign a purchase order for some piece of equipment worth $10,000 without having someone else in the organization approve the purchase and cosign the order?" No more than a half dozen in the crowd of more than a hundred raised their hands. He then asked, "How many people in this room can organize a series of meetings of fifteen or twenty people without having anyone approve the meetings?" The room was full of hands.

The speaker then prodded those in the group, "Is there any difference?" Of course, there is one obvious difference. The purchase order involves *money*; the meetings involve *time*. But every business person (and professor) understands and appreciates the old aphorism, "*Time IS money*." Nevertheless, people everywhere all too often seem to forget that truism when it comes to meetings -- which is understandable, given that the costs of meetings are rarely computed, and when considered are "externalized" (or imposed on others).

Again, we submit that the problem with boring meetings is the incentive structure in the committees. The person calling the meeting will, however, consider the question of whether the meeting is worth his or her own time cost, apart from the costs suffered by others, but notice that the cost suffered by one person is only a minor part of the total cost, and the greater the number of attendees at the meetings, the greater the cost.

The committee problem is similar to the problem of pollution considered at several points in this book because the meeting organizer may determine whether to call a meeting based on some rough comparison between the costs *he* or *she* incurs (but not all committee members incur) and the benefits *he* or *she* receives. However, since the organizer does not incur all of the costs, the meeting is called when there may be few benefits. However, others, following the same logic, also call meetings, the net effect of which is that there can be too many meetings with many of them lasting longer than their economics (the costs and benefits) justify.

Also, at the meetings, every person there may want the meeting to be short and productive, with every comment well thought out and to the point (just as every polluter may want a pond with no detectable waste in it). However, once in the meeting, each committee member can also think like the polluter, "If I make my comment, the meeting will not be extended for long. And the cost to me of my comments is surely lower than the benefits to everyone else hearing my golden words. Besides, if I don't talk, then someone else will. The meeting will be no shorter if I hold back." If everyone thinks that way, then the meeting can easily be consumed with frivolous comments (or "comment pollution"), and meeting length can seem interminable -- or, more accurately, far too long (given the total costs and benefits to *everyone* for the issues considered and the comments made).

This does not mean that all meetings are completely worthless. Meetings do accomplish something of value (or else, we should think, no meetings would ever be called). The problem is that there is an incentive for people, when considering only the costs and benefits of their own situations (their willingness to shirk their duty to restrain themselves and to engage in opportunism), to diminish the meetings' net value by making a "good thing" go on for too long.

What's "too long"? It is when the additional value of a comment made on an additional issue resulting in an additional minute spent in the meeting is less than the cost to all involved. "Too long" means that everyone there would pay all others to keep their mouths shut -- if they could somehow organize themselves to do just that.

Notice that the problem of overly long meetings will likely increase with the number of people in the meeting. This is because the cost an individual incurs when making a comment, which is what the individual can be expected to focus on, stays more or less constant, regardless of how many people join the meeting. However, the total cost to the group -- the "social cost" -- escalates as more members join the meeting. There are simply more people to throw more "waste" into the meeting, with a greater likelihood of the meeting being overly long -- and boring and even unproductive, given that many people may decide to tune out.

As the number of committee members escalates, each member can reason that the decisiveness of his or her votes and comments in affecting committee decisions can diminish. As a consequence, each can conclude that there is less reason to prepare for the meetings, which can mean that comments made may be less well grounded in facts and less well thought out. Each person in a very large meeting may think, "Well, heck, my voice and vote will not affect the outcome of the meeting, so why should I prepare?"

We would be the first to admit that our arguments press the limits of economic reasoning in that we have implicitly assumed that many people in meetings are never considerate of others, and never try to assess the costs of the meetings they call or the comments they make in terms of their impact on others. We recognize that people, at times and to a degree, consider the feelings and costs that they may impose on others. We talk in terms of the logic of the extreme individualist because some people, in and out of business and in and out of meetings, no doubt will think that way. They simply don't consider the costs to others.

However, we also lay out the logic of people consumed by their own private interests because it reveals a force that will be at play even on people who are considerate of others. That force can grow as the committee size grows, neutralizing, at some point, their best intentions and leading to some of the perverse consequences developed from highly strident thinking. Again, we suggest that managers must consider how people will behave in the extreme, not because that is the way everyone behaves all the time in every situation, but because self-serving actions are the type of behavior many human beings exhibit and from which managers must protect themselves through appropriate organizational structures and policies.

More directly to the point, we suggest managers consider our way of thinking because it leads to suggestions for improving the performance of all meetings and committees:

- First, managers ought to find ways of making sure that people who call meetings consider the cost of all involved. We cannot make concrete suggestions, because that requires knowledge of the details of particular work environments. What we do know is that potential committee members have an interest in managers who are tough on the issue of meetings, who are willing to call people to task for unproductive and overly long meetings. Someone, in other words, needs to take charge.
- Second, managers should appoint tough people as chairpersons. These are people who should be willing to cut others off when it is clear they are unprepared and are just sounding off. Managers should recognize that while individuals might prefer meetings in which they can say what they please for as long as they want at the same time everyone else is constrained, the group can still have an interest in tight controls on every member. People are willing to give up some of their own freedom to sound off *if everyone else will, too*.
- Third, managers should be careful about organizing "large" meetings. The productivity of meetings tends to go down as the group size goes up. As a general rule, "small" groups should be organized when *action* is required. "Large" groups should be assembled for the purpose of *reaction* to proposals that have been devised by much smaller groups. If a large committee has been formed and little progress has been made, then the committee should be broken down into smaller working groups, with each subcommittee given a specific assignment that can be presented to the larger committee for final action.
- Fourth, on the other hand, if managers want to give people some sense of participation in the decision-making process without enabling them to actually do anything, then they should make the meetings as large as possible. The participants can be expected to talk without any decisive end, leaving the person who organized the meeting with the authority to take action when something needs to be done.

We suspect that business people are more constrained in meetings than faculty members are by a six-letter word: *Profit*. The goals of a university education are far less

clear, far more elusive and imprecise, because they cannot be relegated to a single bottom-line figure (a fact that, because it works to their advantage, is nurtured by professors). Universities are organized to produce "educated people," which covers a multitude of virtues and sins. This means that the performance of people on committees is hard to assess, and many meetings get bogged down in wrestling with the reasons for the meetings in the first place, with competing factions seeking to elevate their own personal goals above the goals for the committee, if not the university. Business people can, with greater ease, ask a very forceful question that tends to focus the committee process, "What does this (or that) action do for the bottom line?"

In addition, university budgets are typically determined by far-removed state legislatures. Unproductive meetings can easily go undetected within the university bureaucracies, and less by legislators who have little incentive to monitor what the universities do at the committee level. The future welfare of people in the decision-making process is unaffected, one way or the other, by what does or does not go on in any particular meeting. There are simply no close-at-hand residual claimants.

Granted, taxpayers can be thought of as residual claimants, given that efficiency improvement in state university committee processes *can* translate into lower taxes, but each taxpayer has precious little incentive to monitor universities. The monitoring costs can easily exceed the benefits that the individual taxpayer can realize from his or her monitoring, and the probability that the monitoring will have an impact on university efficiency is very close to zero. As we have explained before, taxpayers are all too often the proverbial "free riders" when it comes to monitoring what governments generally do. And when most taxpayers attempt to free ride, they end up getting taken for a ride.

In many regards, faculty members who believe expelling hot air is a virtue can thank their lucky stars for *rationally ignorant taxpayers*. People in business must worry that wasteful meetings will affect their jobs and livelihoods. If firms hold too many meetings, and the bottom line is materially affected, some wise investors will do what cannot be done with universities; the investors will buy the company, eliminate the unproductive meetings, increase the bottom line, and sell the reinvigorated company at a price higher than the purchase price to someone who, because of the price paid, will have an incentive to control meetings.

* * * * *

Managers often spend much of their waking hours trying to figure out how they can make more money by selling more of their product. The lesson to remember is that they can also make money by adjusting their internal structures to account for the impact of numbers on incentives. In short, more than what is produced counts to a firm. Relatively small teams have become increasingly important to business for a number of reasons, but the most important reason is that small teams are a means by which the actions of individual members become meaningful and more easily monitored by others. Teams are a means of discouraging free riding and encouraging everyone to contribute to the value of the whole. Teams are self-enforcing units. Business people would be well advised to apply the principles of teams to the organization of committees.

Concluding Comments

Economists recognize that such considerations as the "importance of the cause" can significantly affect the willingness of the group members to cohere and pursue the common interest of the membership. However, we have concentrated on "large" and "small" groups to demonstrate that, given other factors, an increase in group size beyond some point can have an adverse effect on the motivation which group members have to pursue their common interest. There is, furthermore, substantial evidence to support this basic conclusion. Several studies have revealed that as far as being able to take action, smaller groups, generally with less than seven or eight members, are more efficient than larger ones.[60] Studies also show that as group size within industry increases, job satisfaction tends to decrease and absentee rates, turnover rates, and the incidence of labor disputes tend to increase.[61]

As Mancur Olson points out, even students of history have noticed a difference in the ability of large and small groups to cohere and survive. Olson provides us with this quote from a book by George Homans:

> At the level of. . . the small group, at the level, that is, of a social unit (no matter by what name we call it) each of whose members can have some first-hand knowledge of each of the others, human society, for many millennia longer than written history, has been able to cohere. . . . they have tended to produce a surplus of the goods that make organization successful.
>
> . . . . ancient Egypt and Mesopotamia were civilizations. So were classical India and China; so was Greco-Roman civilization, and so is our own Eastern civilization that few out of medieval Christendom. . . .
>
> the appalling fact is that, after flourishing for a span of time, every civilization but one has collapsed. . . . formal organizations that articulate the whole have fallen to pieces. . . . much of the technology has even been forgotten for lack of the large scale cooperation that could put it in effect. . . . the civilization has slowly sunk to a Dark Age, a situation, much like the one from which it started on its upward path, in which the mutual hostility of small groups is the condition of internal cohesion of each one. . . . Society can fall thus far, but apparently no farther. . . . One can read the dismal story eloquently told, in the historians of civilization from Spengler to Toynbee. The one civilization that has not entirely gone to pieces is our Western Civilization, and we are desperately anxious about it.

---

[60] See, for example, A. Paul Hare, "A Study of Interaction and Consensus in Different-Sized Groups," American Sociological Review, vol. 17, pp. 261-268, June 1952; and John James, "A Preliminary Study of the Size Determinants in Small-Group Interaction," *American Sociological Review*, vol. 16, pp. 444-474, August 1951.

[61] L.W. Porter and EE Lawyer, "Properties of Organization Structure in Relation to Job Attitudes and Job Behavior," *Psychological Bulletin*, 1965, pp. 23-51.

> But at the level of the tribe or group society has always found itself able to cohere.[62]

With a reasonable degree of clarity both the theory and the evidence suggest that if society wishes to pursue some interest that is common to people on a very broad scale, some means other than voluntary group cooperation must be found. It is for this reason we begin again a study of the markets after we look inside the "firm" to see how the "logic of group behavior" explains incentives within firms.

Review Questions

1. Explain why the "free-rider" problem is likely to be greater in a "large" group than in a "small" group.
2. The common interest of people who are in a burning theater is to walk out orderly and in other ways avoid a panic. If that is the case, why do people so frequently panic in such situations? Use rational behavior and the logic of collective action in your answer.
3. Relating to Table5.2, we wrote that Harry is unwilling to eliminate more than Q_1 dandelions and that Fred must bear a portion of the cost of eradicating dandelions if more than Q_1 dandelions are to be eradicated. Explain these statements in terms of the graph.
4. Discuss the costs of making collective decisions in large and small groups. What do these costs have to do with the viability of large and small groups?
5. Intelligent collective decisions can be a common interest shared by members of a large group. Does the analysis of in this chapter suggest anything about the incentive that individuals have to obtain information or about the intelligence of decisions that a large group will make?
6. In what ways do firms overcome the problems discussed in this chapter relating to large groups? How do market pressures affect firm incentives to overcome these problems?
7. Would you expect private firms or government bureaucracies to be more efficient in pursuing the stated "common objectives" of the organization? Explain in terms of the logic of collective action and market forces.
8. You may have a class in which the professor grades according to a curve, whereby he adjusts his or her grading scale to fit the test results. This may also be a class in which everyone in it would prefer not to learn as much as they will. If you are in such a situation (or can imagine one like it), the "common interest " of the class members can be for everyone to study less. The same grading distribution can be

---

[62] George C. Homans, *The Human Group* (New York: Harcourt, Brace, Inc., 1950), pp. 454-456, as cited in Olson, *Logic of Collective Action*, p. 56.

obtained, and everyone can receive his same relative grade for less effort. Why do class members not collude and restrict the amount of studying they do? Would you expect collusion to not study more likely in undergraduate general education courses, core classes in your MBA program, or elective classes in your MBA program?

CHAPTER 6

Reasons for Firm Incentives

Amazing things happen when people take responsibility for everything themselves. The results are quite different, and at times people are unrecognizable. Work changes and attitudes to it, too.

Mikhail Gorbachev

In conventional economic discussions of how firms are managed, incentives are nowhere considered. This is the case because the "firm" is little more than a theoretical "black box" in which things happen somewhat mysteriously. Economists typically acknowledge that the "firm" is the basic production unit, but little or nothing is said of why the firm ever came into existence or, for that matter, what the firm is. As a consequence, we are told little about why firms do what they do (and don't do). There is nothing in conventional discussions that tells us about the role of real people in a firm.

How are firms to be distinguished from the markets they inhabit, especially in terms of the incentives people in firms and markets face? That question is seldom addressed (other than, perhaps, specifying that firms can be one of several legal forms, for example, proprietorships, partnerships, professional associations, or corporations). In conventional discussions of the "theory of the firm," firms maximize their profits, which is their only noted *raison d'être*. But students of conventional theory are never told how firms do what they are supposed to do, or why they do what they do. The owners, presumably, devise ways to ensure that everyone in the organization follows instructions, all of which are intent on squeezing every ounce of profit from every opportunity. Students are never told what the instructions are or what is done to ensure that workers follow them. The structure of incentives inside the firm never comes up because their purpose is effectively assumed away: people do what they are supposed to do, naturally or by some unspecified mysterious process. For people in business, the economist's approach to the "firm" must appear strange indeed, given that business people spend much of their working day trying to coax people to do what they are supposed to do. Nothing is less automatic in business than getting people to pay attention to their firms' profits (as distinguished from the workers' more personal concerns).

In this chapter, before we delve into the structure of firm costs in following chapters, we address the issue of why firms exist not because it is an interesting philosophical question. Rather, we are concerned with that question because its answer can help us understand why the existence of firms and incentives go hand in hand. There is more than an ounce of truth to the refrain, "You cannot have one without the other." In this chapter, we lay out the limited

economic propositions that will undergird the analysis of much of the book. These propositions are powerful as they are simple, are relatively easily to understand.

How Firms Make Markets More Efficient

Why is it that firms add to the efficiency of the markets? That's an intriguing question, especially given how standard theories trumpet the superior efficiency of markets. Students of conventional theory might rightfully wonder: If markets are so efficient, why do entrepreneurs ever go to the trouble of organizing firms? Why not just have everything done by way of markets, with little or nothing actually done (in the sense that things are "made") inside firms? All of the firm's inputs could be bought by individuals, with each individual adding value to the inputs he or she purchases and then selling this result to another individual who adds more value, etc. until a final product is produced and a final market is reached at which point the completed product is sold to consumers. The various independent suppliers may be at the same general location, even in the same building, but everyone, at all times, could be up for contracting with all other suppliers or some centralized buyer of the inputs. By keeping everything on a market basis, the benefits of competition could be constantly reaped. Entrepreneurs could always look for competitive bids from alternative suppliers for everything used -- whether in the form of parts to be assembled, accounting and computer services to be used, or, for that matter, executive talent to be employed.

Individuals, as producers relying exclusively on markets, could always take the least costly bid. They could also keep their options open, including retaining the option to switch to new suppliers that propose better deals. No one would be tied down to internal sources of supply for their production needs. They would not have to incur the considerable costs of organizing themselves into production teams and departments and various levels of management. They would not have to incur the costs of internal management. They could, so to speak, maintain a great deal of freedom!

Then why do firms exist? What is the incentive – driving force – behind firms? For that matter, what is a *firm* in the first place? University of Chicago Law and Economics Professor Ronald Coase, on whose classic work "The Nature of the Firm" much of this chapter is based and many of the particular arguments drawn, proposed a substantially new but deceptively simple explanation.[1] He reasoned that the *firm* is any organization that supercedes the pricing system, in which hierarchy, and methods of command and control are substituted for exchanges. To use his exact words: "A firm, therefore, consists of the system of relationships which comes into existence when the direction of resources is dependent on an entrepreneur."[2]

---

[1] Ronald H. Coase, "The Nature of the Firm," Economica, vol. 4 (1937), pp. 386-405, reprinted in R. H. Coase, The Firm, the Market, and the Law (Chicago: University of Chicago Press, 1988), pp. 33-55.
[2] Ibid., pp. 41-42.

Good answers to the question of why firms exist is more complicated and longer in the making than might be thought, but space limitations of this book require us to be brief. Some economists have speculated that firms exist because of the *economies of specialization* of resources, a key one being labor. Clearly, Adam Smith and many of his followers were correct when they observed that when tasks are divided among a number of workers, the workers become more proficient at what they do. Smith began his economic classic The Wealth of Nations by writing about how specialization of labor increased "pin" (really nail) production.[3] By specializing, workers can become more proficient at what they do, which means they can produce more in their time at work. They also don't have to waste time changing tasks, which means more time can be spent directly on production.

While efficiency improvements can certainly be had from specialization of any resource, especially labor, Smith was wrong to conclude that firms were *necessary* to coordinate the workers' separate tasks. This is because, as economists have long recognized, their separate tasks could be coordinated by the pricing system within markets.

Markets could, conceivably, exist even within the stages of production that are held together by, say, assembly lines. Workers at the various stages could simply buy what is produced before them. The person who produces soles in a shoe factory could buy the leather and then sell the completed soles to the shoe assemblers. For example, the bookkeeping services provided a shoe factory by its accounting department could easily be bought on the market. Similarly, all of the intermediate goods involved in Smith's pin production could be bought and sold until the completed pins are sold to those who want them.

Why, then, do we observe *firms* as such, which organize activities by hierarchies and directions that are not based on changing prices (which distinguishes them from markets)? In terms of our examples, why are there shoe and pin companies? Admittedly, over the years economists have tendered various answers.[4]

---

[3] Adam Smith, An Inquiry into the Nature and Causes of the Wealth of Nations (New York: Modern Library, 1937), pp. 4-12.

[4] The late University of Chicago economist Frank Knight speculated that firms arise because of *uncertainty* (Risk, Uncertainty, and Profit [Chicago: University of Chicago Press, 1971]). If business were conducted in a totally certain world, there would be no need for firms, according to Knight. Workers would know their pattern of rewards, and there would be no need for anyone to specialize in the acceptance of the costs of dealing with risks and uncertainties that abound in the real world of business.

As it is, according to Knight, some workers are willing to work for firms because of the type of deal that is struck: The workers accept a reduction in their expected pay in order to reduce the variability and outright uncertainty of that pay. Entrepreneurs are willing to make such a bargain with their workers because they are effectively paid to do so by their workers (who accept a reduction in pay) and because the employers can reduce their exposure to risk and uncertainties faced by individual workers by making similar bargains with a host of workers. As Knight put it (see the bottom of the next page),

> This fact [the intelligence of one person can be used to direct others] is responsible for the most fundamental change of all in the form of organization, the system under which the confident and adventuresome assume the risk or insure the doubtful and timid by

Again, how can the existence of *firms*, as constructs distinctly different from markets, be explained? There are probably many reasons people might think firms exist, several of which Coase dismisses for being wrongheaded or for not being important.[5] What Coase was interested in, however, was not a catalogue of "small" explanations for this or that firm, but an explanation for the existence of firms that, to one degree or another, is applicable to virtually all firms. He was seeking a unifying theme, a common basis. In his 1937 article, he struck upon an unbelievably simple answer to his puzzle, but it was also an explanation that earned him the Nobel Prize in Economics -- more than a half-century later!

What did he say? How did he justify the firm's existence? Simply put, he observed that there are costs of dealing in markets. He dubbed these costs *marketing costs*, but most economists now call them *transaction costs.* Whatever they are called, these costs include the time and resources that must be devoted to organizing economic activity through markets. Transaction costs include the particular real economic costs (whether measured in money or not) of discovering the best deals as evaluated in terms of prices and attributes of products, negotiating contracts, and ensuring that the resulting terms of the contract are followed. When we were going through our explanation of how work on an assembly line could be viewed as passing through various markets, most readers probably imagined that the whole process could be terribly time consuming, especially if the suppliers and producers at the various stages were constantly subject to replacement by competitors.

Reasons Firms Exist

Once the costs of market activity are recognized, the reason for the emergence of the firm is transparent: *Firms, which substitute internal direction for markets, arise because they reduce the need for making market transactions. Firms lower the costs that go with market transactions.* If internal direction were not, at times and up to some point, more cost-

---

> guaranteeing to the latter a specified income in return for the assignment of the actual results . . . With human nature as we know it, it would be impracticable or very unusual for one man to guarantee to another a definite result of the latter's actions without being given power to direct his work. And on the other hand the second party would not place himself under the direction of the first without such a guarantee . . . The result of this manifold specialization of function is the enterprise and wage system of industry. Its existence in the world is the direct result of the fact of uncertainty (Ibid., pp. 269-270).

[5] Ibid, pp. 41-42. For example, Coase concedes that some people might prefer to be directed in their work. As a consequence, they might accept lower pay just to be told what to do. However, Coase dismisses this explanation as unlikely to be important because "it would rather seem that the opposite tendency is operating if one judges from the stress normally laid on the advantage of 'being one's own master'" (Ibid., p. 38). Of course, it might be that some people like to control others, meaning they would give up a portion of their pay to have other people follow their direction. However, again Coase finds such an explanation lacking, mainly because it could not possibly be true "in the majority of the cases." (Ibid.). People who direct the work of others are frequently paid a premium for their efforts.

effective than markets, then firms would never exist – would have no reason for being, meaning that no one would have the required incentive to go to the trouble of creating them. However, while firms may never eliminate the need for markets and contracts, with all of their attendant costs, they must surely reduce them.

Entrepreneurs and their hired workers essentially substitute one long-term contract for a series of short-term contracts: The workers agree to accept directions from the entrepreneurs (or their agents, or managers) within certain broad limits (with the exact limits subject to variation) in exchange for security and a level of welfare (including pay) that is higher than the workers would be able to receive in the market without firms. Similarly, the entrepreneurs (or their agents) agree to share with the workers some of the efficiency gains obtained from reducing transaction costs.[6]

The firm is a viable economic institution because both sides to the contract – owners and workers -- gain. Firms can be expected to proliferate in markets simply because of the mutually beneficial deals that can be made. Those entrepreneurs who refuse to operate within firms and stick solely to market-based contracts, when in fact a firm's hierarchical organization is more cost-effective than market-based organizations, will simply be out-competed for resources by the firms that do form and achieve the efficiency-improving deals with workers (and owners of other resources).

If firms reduce transaction costs, does it follow that one giant firm should span the entire economy, as, say, Lenin and his followers thought possible for the Soviet Union? Our intuition says, "No!" But there are also good reasons for expecting firms to be limited in size.

Cost Limits to Firm Size

Clearly, by organizing activities under the umbrella of firms, entrepreneurs give up some of the benefits of markets, which provide competitively delivered goods and services. Managers suffer from their own limited organizational skills, and skilled managers are scarce, as evident by the relatively high salaries many of them command. Communication problems within firms expand as firms grow, encompassing more activities, more levels of production, and more diverse products. Because many people may not like to take directions, as the firm expands to include more people, the firm may have to pay progressively higher prices to workers and other resource owners in order to draw them into the firm and then direct them.

---

[6]Coase recognizes that entrepreneurs could overcome some of the costs of repeatedly negotiating and enforcing short-term contracts by devising one long-term contract. However, as the time period over which a contract is in force is extended, more and more unknowns are covered, which implies that the contract must allow for progressively greater flexibility for the parties to the contract. The firm is, in essence, a substitute for such a long-term contract in that it covers an indefinite future and provides for flexibility. That is to say, the firm as a legal institution permits workers to exit more or less at will and it gives managers the authority, within bounds, to change the directives given to workers.

There are, in short, limits to what can be done through organizations. These limits can't always be overcome, except at greater costs, even with the application of the best organizational techniques, whether through the establishment of teams, through the empowerment of employees, or through the creation of new business and departmental structures (for example, relying on top-down, bottom-up, or participatory decision making). Even the best industrial psychology theories and practices have their limits when applied to human relationships.

The Agency Problem

Firms might be restricted in their size because they are also likely to suffer from a major problem -- the so-called *agency problem* (or, alternately, the *principal/agent problem*) that will be considered and reconsidered often in this book. This problem is easily understood as a conflict of interests between identifiable groups within firms. The entrepreneurs or owners of firms (the *principals*) organize firms to pursue their own interests, which are often (but, admittedly, not always) greater profits. To pursue profits, however, the entrepreneurs must hire managers who then hire workers (all of whom are *agents*). However, the goals of the worker/agents are not always compatible with the goals of the owner/principals. Indeed, they are often in direct conflict. Both groups want to get as much as they can from the resources assembled in the firms.

The problem the principals face is getting the agents to work diligently at their behest and with their (the principals') interests in mind, a core problem facing business organizations that even the venerable Adam Smith recognized more than two centuries ago.[7] Needless to say, agents often resist doing the principals' bidding, a fact that makes it difficult -- i.e., costly -- for the principals to achieve their goals.

It might be thought that most, if not all, of these conflicts can be resolved through contracts, which many can. However, like all business arrangements, contracts have serious limitations, not the least of which is that they can't be all-inclusive, covering all aspects of even "simple" business relationships (which all are more or less complex). Contracts simply cannot anticipate and cover all possible ways the parties to the contract, if they are so inclined, can get around specific provisions. The cost of enforcing the contracts can also be a problem, and an added cost, even when both parties know that provisions have been violated. Each party will recognize the costs and may be tempted to exploit them, and will figure that the other may be

[7] In his classic The Wealth of Nations, Adam Smith wrote, "The directors of such companies, however, being the managers rather of other people's money than of their own, it cannot be well expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own. Like the stewards of a rich man, they are apt to consider attention to small matters as not for their master's honour, and very easily give themselves a dispensation from having it. Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company" [The Wealth of Nations (New York, Modern Library, 1937), p. 700].

equally tempted. Each will seek some means by which the contract will be self-enforcing, or will encourage each party to live up to the letter and spirit of the contract because it is in the interest of each party to do so. This is where incentives will come in, to help make contracts self-enforcing. Incentives can encourage the parties to more closely follow the intent and letter of contracts.

Competition will be a powerful force toward minimizing agency costs. Firms in competitive markets that are not able to control agency costs are firms that are not likely to survive for long, mainly because of what has been dubbed the "market for corporate control."[8] Firms that allow agency costs to get out of hand will risk either failure or takeover (by way of proxy fights, tender offers, or mergers). In later chapters, we will discuss at length how managers can solve their own agency problems including controlling their own behavior as agents for shareholders. At the same time, we would be remiss if we didn't repeatedly point out the market pressures on managers to solve such problems, even if they are not naturally inclined to do so. If corporations are not able to adequately solve their agency problems, we can imagine that the corporate form of doing business will be (according to one esteemed financial analyst[9]) "eclipsed" as new forms of business emerge. Of course, this means that obstruction in the market for corporate control (for example, legal impediments to takeovers) can translate into greater agency costs, and less efficient corporate governance.

Why are firms the sizes they are? When economists in or out of business usually address that question, the answer most often given relates in one way or another to *economies of scale*. By economies of scale, we mean something very specific, the cost savings that emerge when all resource inputs -- labor, land, and capital -- are increased together. In some industries, it is indeed true that as more and more of all resources are added to production within a given firm, output expands by more than the use of resources. That is to say, if resource use expands by 10 percent and output expands by 15 percent, then the firm experiences economies of scale. Its (long-run) average cost of production declines. Why does that happen? The answer is almost always "technology," which is another way of saying that it "just happens," given what is known about combining inputs and getting output. This is not the most satisfying explanation, but it is nonetheless true that economies of scale are available in some industries (automobile) but not in others (crafts).

We agree that the standard approach toward explaining firm size is instructive. We have spent long hours at our classroom boards with chalk in hand developing and describing scale economies in the typical fashion of professors, using (long-run) average cost curves and pointing out when firms in the expansion process contemplate starting a new plant. We think the

---

[8] One of the more important contemporary articles on the "market for corporate control" is by Henry G. Manne, "Mergers and the Market for Corporate Control," Journal of Political Economy, vol. 73 (April 1963), pp. 110-120.

[9] See Michael C. Jensen, "Eclipse of the Public Corporation," Harvard Business Review (September-October 1989), pp. 64-65.

standard approach is useful, but we also believe it leaves out a lot of interesting forces at work on managers within firms. This is understandable, given that standard economic theory totally assumes away the roles of managers, which we intend to discuss at length.

Coase and his followers have taken a dramatically different tack in explaining why firms are the sizes they are in terms of scale of operations and scope of products delivered to market. The new breed of theorists pays special attention to the difficulties managers face as they seek to expand the scale and scope of the firm. They posit that as a firm expands, agency costs mount. This is primarily because workers have more and more opportunities to engage in what can only be tagged *opportunistic behavior* – or taking advantage of their position by misusing and abusing firm resources. *Shirking*, or not working with due diligence, is one form of opportunistic behavior that is known to all employees. Theft of firm resources is another form. As the firm grows, the contributions of the individual worker become less detectable, which means workers have progressively fewer incentives to work diligently on behalf of firm objectives, or to do what they are told by their superiors. They can more easily hide.

The tendency for larger size to undercut the incentives of participants in any group is not just theoretical speculation. It has been observed in closely monitored experiments. In an experiment conducted more than a half century ago, a German scientist asked workers to pull on a rope connected to a meter that would measure the effort expended. Total effort for all workers combined increased as workers were added to the group doing the pulling at the same time that the individual efforts of the workers declined. When three workers pulled on the rope, the individual effort averaged 84 percent of the effort expended by one worker. With eight workers pulling, the average individual effort was one-half the effort of the one worker.[10] Hence, group size and individual effort were inversely related – as they are in most group circumstances -- inversely related.

The problem evident in the experiment is not that the workers become any more corrupt or inclined to take advantage of their situation as their number increases. The problem is that their incentive to expend effort deteriorates as the group expands. Each person's effort counts for less in the context of the larger group, a point which University of Maryland economist Mancur Olson elaborated upon decades ago (and we considered in detail in the last chapter).[11] The "common objectives" of the group become less and less compelling in directing individual

---

[10] As reported by A. Furnham, "Wasting Time in the Board Room," Financial Times, March 10, 1993, p. xx.

[11] See Mancur Olson, The Logic of Collective Action: Public Goods and the Theory of Groups (Cambridge, Mass.: Harvard University Press, 1965). Olson argues that common goals have less force in "large" groups than "small" groups, which explains why cartels don't form in open competitive markets. All competitors might understand that it is in their group interest to cut production and increase their market price, if all curb production. However, each competitor can reason that its individual curb in output will have no effect on total output and thus cannot be detected. Hence, the "logic of collective action" is for everyone to "cheat" on the cartel, or not curb production, which means that nothing will happen to the market price.

efforts. Such a finding means that if each worker added to the group must be paid the same as all others, the cost of additional production obviously rises with the size of the working group. The finding also implies that to get a constant increase in effort with the additional workers, all workers must be given greater incentive to hold to their previous level of effort.[12]

Optimum Size Firms

How large should a firm be? Contrary to what might be thought, the answer depends on more than "economies of scale" technically specified. Technology determines what *might* be possible, but it doesn't determine what *will* happen. And what happens depends on policies that minimize shirking and maximize the use of the technology by workers. This means that scale economies depend as much or more on what happens within any given firm as they do on what is technologically possible. The size of the firm obviously depends on the extent to which owners must incur greater monitoring costs as they lose control with increases in the size of the firm and additional layers of hierarchy (a point well developed by Oliver Williamson in his classic article written more than thirty years ago[13]). However, the size of the firm also depends on the cost of using the market.

Management information Professors Vijay Gurbaxani and Seungjin Whang have devised a graphical means of illustrating the "optimal firm size" as the consequence of two forces: "internal coordinating costs" and "external coordinating costs."[14] As a firm expands, its internal coordinating costs are likely to increase. This is because the firm's hierarchical pyramid will likely become larger with more and more decisions made at the top by managers who are further and further removed from the local information available to workers at the bottom of the pyramid. There is a need to process information up and down the pyramid. When the information goes up, there are unavoidable problems and costs: costs of communication, costs of miscommunication, and opportunity costs associated with delays in communication, all of which can lead to suboptimal decisions. These "decision information costs" become progressively greater as the decision rights are moved up the pyramid.

Attempts to rectify the decision costs by delegating decision making to the lower ranks may help, but this can – and *will* -- also introduce another form of costs -- which, you will recall, we previously have called *agency costs.* These include the cost of monitoring (managers

---

[12] Workers can also reason that if the residual from their added effort goes to the firm owners, they can possibly garner some of the residual by collusively (by explicit or tacit means) restricting their effort and hiking their rate of pay, which means that the incentive system must seek to undermine such collusive agreement. For a discussion of these points see, Felix R. FitzRoy and Kornelius Kraft, "Cooperation, Productivity, and Profit Sharing," Quarterly Journal of Economics (February 1987), pp. 23-35.

[13] Oliver E. Williamson, "Hierarchical Control and Optimum Size Firms," Journal of Political Economy , vol. 75 (no. 2, 1967), pp. 123-138.

[14]Vijay Gurbaxani and Seungjin Whang, "The Impact of Information Systems on Organizations and Markets," Communication of the ACM, January 1991, pp. 59-73.

actually watching employees as they work or checking their production) and bonding (workers providing assurance that the tasks or services will be done as the agreement requires), and the loss of the residual gains (or profits) through worker shirking, which we covered earlier.

The basic problem managers face is one of balancing the decision information costs with agency costs and finding that location for decision rights that minimizes the two forms of costs. From this perspective, where the decision rights are located will depend heavily on the amount of information flow per unit of time. When upward flow of information is high, the decision rights will tend to be located toward the floor of the firm, mainly because the costs of suboptimal decisions by having the decision making done high up the hierarchy will be high. The firm, in other words, can afford to tolerate agency costs because the costs of avoiding the them, via centralized decisions, can be higher.

Nevertheless, as the firm expands, we should expect that the internal coordinating costs along with the cost of operations will increase. The upward sloping line in Figure 6.1 depicts this relationship.

But internal costs are not all that matter to a firm contemplating an expansion. It must also consider the cost of the market, or what Gurbaxani and Whang call "external coordination costs." If the firm remains "small" and buys many of its parts, supplies, and services (such as accounting, legal, and advertising services) from outside venders, then it must cover a number of what we have called "transaction costs." These include the costs of transportation, inventory holding, communication, contract writing, and contract enforcing. However, as the firm expands in size, then these transaction costs should be expected to diminish. After all, a larger firm seeks to supplant market transactions. The downward sloping line in Figure 6.1A depicts this inverse relationship between firm size and transaction costs.

Again, how large should a firm be? If a firm vertically integrates, it will engage in fewer market transactions, lowering its transaction costs. It can also benefit from economies of scale, the technical kind mentioned earlier. However, in the process of expanding, it will confront growing internal coordination costs, or all of the problems of trying to move information up the decision making chain, getting the "right" decisions, and then preventing people from exploiting their decision making authority to their own advantage.

The firm should stop expanding in scale and scope when the total of the two types of costs -- external and internal coordinating costs -- are minimized. This minimum can be shown graphically by summing the two curves in Figure 6.1A to obtain the U-shaped curve in Figure 6.1B. The *optimal* (or most efficient/cost-effective) firm size is at the bottom of the U.

This way of thinking about firm size would have only limited interest if it did not lend itself to a couple of additional observations, which permit thinking about the location, shape, and changes in the curve. First, the exact location of the bottom will, of course, vary for different firms in different industries. Different firms have different capacities to coordinate activities

through markets and hierarchies. Second, firm size will also vary according to the changing abilities of firms to coordinate activities internally and externally.

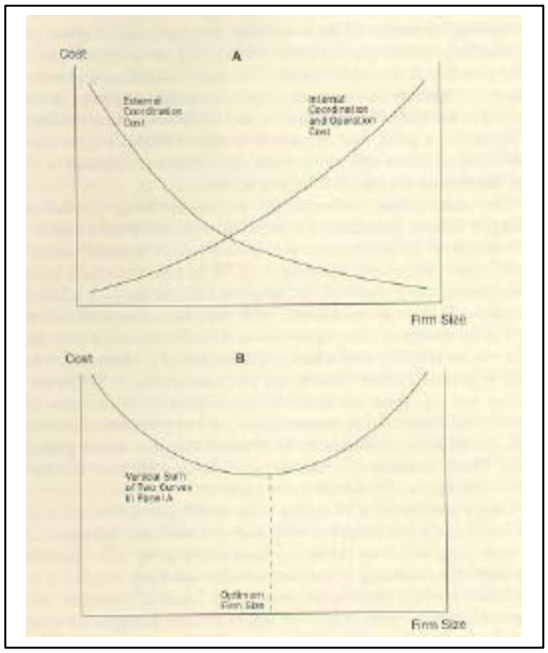


Figure 6.1A and 6.1B External and Internal Coordinating Costs
As the firm expands, the internal coordinating costs increase as the external coordinating costs fall. The optimum firm size is determined by summing these two cost structures, which is done in the bottom half of the figure.

A firm that is efficient at processing information will be larger, everything else equal, than one that isn't so able. If a firm is able to improve the efficiency of its upward information flow and reduce the number of wrong decisions, then the upward sloping curve in Figure 6.1A will move down and to the right, causing the sum of the two curves in the bottom panel of the figure to move to the right, for a greater optimal size firm. If the costs of using markets go down, the firm size can be expected to decline, not because the firm has become less efficient internally (it may have become more efficient), but because markets are now relatively more cost effective. Again, from this perspective, the size of the firm changes for reasons other than those related to the technology of actual production. It depends on the ability of managers to squeeze out the scale economies that are possible from their workers.

Of course, knowing that the owners will always worry that their manager-agents will exploit their positions for their own benefit at the expense of the owners, managers will want to "bond" themselves against exploitation of their positions. (And we don't use the term "bond" in the modern pop-psychology sense of developing warm and fuzzy relationships; rather, we use it in the same sense that is common when accused criminals post a bond, or give some assurance that they will appear in court if released from jail.) That is to say, managers have an interest in letting the owners know that they, the managers, will suffer some loss when exploitation occurs.

Devices such as audits of the company are clearly in the interest of stockholders. But they are also in the interest of managers by reducing the scope for managerial misdeeds, thus increasing the market value of the company – and the value of its managers. By buying their companies' stock, manager-agents can also bond themselves, assuring stockholders that they will incur at least some losses from agency costs. To the extent manager-agents can bond themselves convincingly, the firm can grow from expanded sources of external investment funds. By bonding themselves, manager-agents can demand higher compensation. Firms can be expected to expand and contract with reductions and increases in the costs of developing effective managerial bonds.[15]

Changes in Organizational Costs

Finally, we can observe that the size of the firm can be expected to change with changes in the relative costs of organizing a given set of activities by way of markets and hierarchies. For example, suppose that the costs of engaging in market transactions are lowered, meaning markets become relatively more economical *vis a vis* firms. Entrepreneurs should be expected to organize more of their activities through markets, fewer through firms. Then, those firms that more fully exploit markets, and rely less on internal directions, should be able to increase the payments provided workers and other resources that they buy through markets, collectively leaving fewer resources to expand their market share relative to those firms that make less use of markets. Accordingly, firms should be expected to *downsize*, to use a popular expression.

An old, well-worn, and widely appreciated explanation for downsizing is that modern technology has enabled firms to produce more with less. Personal computers, with their ever-escalating power, have enabled firms to lay off workers (or hire fewer workers). Banks no longer need as many tellers, given the advent of the ATMs.

One not-so-widely-appreciated explanation is that markets have become cheaper, which means that firms have less incentive to use hierarchical structures and more incentive to use markets. And one good reason firms have found markets relatively more attractive is the rapidly developing computer and communication technology, which has reduced the costs of entrepreneurs operating in markets. The new technology has lowered the costs of locating suitable trading partners and suppliers, as well as negotiating, consummating, and monitoring market-based deals (and the contracts that go with them). In terms of Figure 6.1, the downward sloping transaction costs curve has dropped down and to the left, causing the bottom of the U to move leftward.

"Outsourcing" became a management buzzword in the 1980s because the growing efficiency of markets, through technology, made it economical. Outsourcing continued apace in the 1990s. Of 26 major companies surveyed, 86 percent said they outsourced some activity in

[15] See Michael C. Jensen and William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," Journal of Financial Economics, vol. 3 (October 1976), pp. 325-328.

1995, up from 58 percent who gave the same response in 1992, with the budding outsourcing industry generating $100 billion in annual revenues by 1996.[16] For all practical purposes, airlines now outsource the acquisition of their reservations through independent contractors called travel agents, given that more than 70 percent of all airline reservations are now taken by such agents, working through computerized markets, not through the hierarchical structures within the airlines.

Modern technology has also improved the monitoring of employees, reducing agency costs, which has been a force for the expansion of firms. This is because firms have been able to use the technology to garner more of the gains from economies of scale and scope. The optical scanners at grocery store checkout counters are valuable because they can speed up the flow of customers through the checkout counters, but they can also be used for other purposes, such as inventory control and restocking. Each sale is immediately transmitted to warehouse computers that determine the daily shipments to stores. The scanners can also be used to monitor the work of the clerks, a factor that can diminish agency costs and increase the size of the firm. (We are told that even "Employee of the Month Awards" are made based on reports from scanners.) Books on Tape, a firm that rents audio versions of books, tracks its production of tapes by way of scanners not so much to reward and punish workers, but to be able to identify problem areas. In terms of Figure 6.1, the upward sloping curve moves down and to the right, while the U-shaped curve in the lower panel moves to the right.

Frito-Lay has issued its sales people hand scanners in part to increase the reliability of the flow of information back to company distribution centers, but also to track the work of the sales people. The company can obtain reports on when each employee starts and stops work, the time spent on trips between stores, and the number of returns. The sales people can be asked to account for more of their time and activities while they are on the job.

Obviously, we have not covered the full spectrum of explanations for the rich variety of sizes of firms that exists in the "real world" of business. We have also left the net impact of technology somewhat up in the air, given that it is pressing some firms to expand and others to downsize. The reason is simple: technology is having a multitude of impacts that can be exploited in different ways by firms in different situations.

Prisoners' Dilemma Problems, Again

The discussion to this point reduces to a relatively simple message: Firms exist to bring about cost savings, and they generate the cost savings through cooperation. However, cooperation is not always and everywhere "natural"; people have an incentive to "cheat," or not do what they are supposed to do or have agreed to do. This may be the case because of powerful incentives to toward noncooperation built-in to many business environments.

---

[16] As reported by John A. Byrne, "Has Outsourcing Gone Too Far?" Business Week, April 1, 1996, p. 27.

An illustration of the tendency toward noncooperative behavior, despite the general advantage from cooperation, is a classic so-called "conditional-sum game" known as the *prisoners' dilemma* (which we have already introduced without formally calling them by their proper name).[17] This is a dilemma, commonly found in business, that takes its name from a particular situation involving the decision two prisoners have to make on whether or not to confess to a crime they committed. But the dilemma can also be applied whenever two or more people find themselves in a situation where the best decision from the perspective of each leads to the worst outcome from the perspective of all.

Consider a situation in which the police have two people in custody who are known to be guilty of a serious crime, but who, in the absence of a confession by one of them, can be convicted only of a relatively minor crime. How can the police (humanely) encourage the needed confession? One effective approach is to separate the two prisoners and present each with the same set of choices and consequences. Each is told that if one confesses to the serious crime and the other does not, then the one who confesses receives a light sentence of one year, while the one who does not confess receives the maximum sentence of fifteen years. If they both confess, then both receive the standard sentence of ten years. And if both refuse to confess, then each is sentenced to two years for the minor crime.

The choices and consequences facing the prisoners are presented in the "payoff" matrix in Table 6.1, where the first number in each parenthesis is the sentence in years received by prisoner A, and the second number is the sentence received by prisoner B:

Table 6.1 Prisoners' Dilemma

		B	
		Don't Confess	**Confess**
A	**Don't Confess**	(2 2)	(15 1)
	Confess	(1 15)	(10 10)

From the perspective of both prisoners the best outcome occurs if neither one confesses (they serve a total of four years), and the worst outcome occurs if both confess (they serve a total of 20 years). In other words, if both prisoners cooperate with each other by keeping their mouths shut, they will both be far better off than if they act noncooperatively with each other by confessing. However, from the perspective of each prisoner the best choice is the noncooperative one of confession.

Consider the situation from prisoner A's vantage point. If A believes that B will refuse to confess, then he receives two years in prison if he also refuses to confess, but only one year if

[17] "Conditional-sum games" are games in which the value available to the participants is dependent how the game is played.

he does confess. His best choice is to confess. On the other hand, if A believes that B will confess, then he receives fifteen years in prison if he does not confess and only ten years if he does. Again, his best choice is to confess. No matter what A believes B will do, it is in A's best interest to confess. And the incentives are exactly the same for B. So while it is rational from their individual perspectives for both A and B to make the noncooperative choice, the result is the worst possible outcome from their collective perspective.

When, as in our example, only two people are in a prisoners' dilemma setting, it is quite possible for them to avoid the worst outcome by choosing the cooperative option of not confessing. The two prisoners may be good friends and have genuine regard for the well being of each other, in which case each will feel confident that the other will not betray him with a confession, and will refuse to betray his friend. But if the number of prisoners grows and becomes quite large, then it becomes much less likely that any one of them can reasonably trust everyone else to keep quiet. This means that as the number grows it becomes increasingly irrational for any one of them to keep quiet.

Overcoming the Large-Numbers Prisoners' Dilemma Problems

Overcoming a large-number prisoners' dilemma by motivating cooperative behavior is obviously difficult, but not impossible. *The best hope for those who are in a prisoners' dilemma situation is to agree ahead of time to certain rules, restrictions, or arrangements that will punish those who choose the noncooperative option.* For example, those who are jointly engaging in criminal activity will see advantages in forming gangs whose members are committed to punishing noncooperative behavior. The gang members who are confronted with the above prisoners' dilemma will seriously consider the possibility that the shorter sentence received for confessing will hasten the time when a far more harsh punishment for "squealing" on a fellow gang member is imposed by the gang.

The problem illustrated by the prisoners' dilemma is a very general one that is encountered in many different guises, most of which have nothing to do with prisoners. Excessive pollution, for example, can be described as a prisoners' dilemma in which citizens – meaning, typically, a very large number of people -- would be better off collectively if everyone polluted less, yet, from the perspective of each individual the greatest payoff comes from continuing to engage in polluting activities no matter what others are expected to do. As another example, while there may be wide agreement that we would be better off with less government spending, each interest group is better off lobbying for more government spending on its favorite program. People are tempted by the noncooperative solution in polluting and lobbying because they benefit individually and only have limited and costly ways of ensuring that others resist the noncooperative solution.

Many areas of business are fertile grounds for the conditional-sum game situations represented by the prisoners' dilemma. A number of examples of business-related prisoners' dilemmas will be discussed in some detail in subsequent chapters, and an important task of managers is to identify and resolve these dilemmas as they arise both within the firm and with suppliers and customers of the firm. Indeed, we see "management" as concerned with finding resolutions of prisoners' dilemmas. Good managers constantly seek to remind members of the firm of the benefits of cooperation and of the costs that can be imposed on people who insist on taking the noncooperative course.

Consider, for example, the issue of corporate travel, which is a major business expense -- estimated at over $130 billion in 1994 (the latest available data at this writing).[18] If a business were able to economize on travel costs, it would realize significant gains. And much of this gain would be captured by the firms' traveling employees who, if they were able to travel at less cost, would earn higher incomes as their net value to the firm increased. So all the traveling employees in a firm could be better off if they all cut back on unnecessary travel expenses. But the employees are in a prisoners' dilemma with respect to reducing travel costs because each recognizes that he or she is personally better off by flying first class, staying at hotels with multiple stars, and dining at elegant restaurants (behaving noncooperatively), than making the least expensive travel plans (behaving cooperatively) regardless of what the other employees do. Each individual employee would be best off if all other employees economized, which would allow her salary to be higher as she continued to take luxury trips. But if the others also make the more expensive travel arrangements, she would be foolish not to do so herself since her sacrifice would not noticeably increase her salary.

Airlines have recognized the "games" people play with their bosses and other workers, and have played along by making the travel game more rewarding to business travelers, more costly to the travelers' firms, and more profitable to the airlines – all through their "frequent-flier" programs. Of course, you can bet managers are more than incidentally concerned about the use of frequent-flier programs by employees. When American Airlines initiated its AAdvantage frequent-flier program in 1981, the company was intent on staving off the fierce price competition that had broken out among established and new airlines after fares and routes were deregulated in 1978. As other writers have noted, American was seeking to enhance "customer loyalty" by offering their best, most regular customers free or reduced-price flights after they built up their mileage accounts. Greater customer loyalty can mean that customers are less responsive to price increases, which could translate into actual higher prices. [19]

At the same time, there is more to the issue than "customer loyalty." American figured that it could benefit from the obvious prisoners' dilemma their customers, especially business

---

[18] As reported in Jonathan Dahl, "Many Bypass the New Rules of the Road," The Wall Street Journal, September 29, 1994, p. B1.

[19] For a discussion of frequent-flier programs as a means of enhancing customer loyalty, see Adam M. Brandenburger and Barry J. Nalebuff, Co-opetition (New York: Currency/Doubleday, 1996), pp. 132-158.

travelers, are in. By setting up the frequent-flier program, American (and all other airlines that followed suit) increased the individual payoff to business travelers for noncooperative behavior. American did this under its frequent-flier program by allowing travelers to benefit from more free flights and first-class upgrades by choosing more expensive, and often less direct, flights. They encouraged business people to act opportunistically, to use their discretion for their own benefit at the expense of everyone else in their firms.

For example, a business traveler who is on the verge of having enough miles in his American account to qualify for elite status (additional upgrades of travel perks) might choose a more expensive American flight over a comparable Southwest Airline flight just to get additional AAdvantage miles. The company would in effect, pick up the cost of the traveler's vacation flight. Business travelers are also encouraged to book their flights later than they could, which requires paying full fare, so they can use their frequent-flier upgrades to first class (these upgrades are typically not allowed with discount tickets). Or business people will take circuitous routes to their destinations to qualify for more frequent-flier miles than could be gotten from a direct trip. The prisoners' dilemma problem for workers and their companies has, of course, prompted as a host of other non-airline firms -- rental car companies, hotels, and restaurants – to begin granting frequent-flier miles with selected airlines for travel services people buy with them, encouraging once again higher-than-necessary travel costs. The company incurs the cost of the added miles plus the lost time.

Now, use of frequent-flier miles might actually lower worker wages (because of the added cost to their firms, which can reduce the demand for workers, and the benefit of the miles to workers, which can increase worker supply and lower wages, topics to be covered later), but, still, workers have an incentive to exploit the program. Again, they are in prisoners' dilemma under which the cooperative strategy might be best for all, but the noncooperative strategy dominates the choice each individual faces.

These problems created by frequent-flier programs are not trivial for many businesses, and we would expect the bigger the firm, the greater the problem (given the greater opportunity for opportunistic behavior in large firms). Thirty percent of business travelers working for Mitsubishi Electronics America wait until the last few days before booking their flights, according to corporate travel manager John Fazio. Fazio adds, "We have people who need to travel at the last minute, but it's not 30 percent."[20] Corporate travel managers complain that the frequent-flier programs have resulted in excessive air fares (a problem for 87 percent of the firms surveyed), wasted employee time (a problem for 68 percent of the surveyed firms), use of more expensive hotels (a problems for 67 percent of the surveyed firms), and unnecessary travel (a problem for 59 percent of the surveyed firms).[21] The corporate travel managers

[20] See Dahl, Ibid., p. B1.

[21] As reported by Frederick J. Stephenson and Richard J. Fox, "Corporate Strategies for Frequent-Flier Programs," Transportation Journal, vol. 32, no. 1, (Fall 1992), pp. 38-50. The 1991 survey included 506 corporate members of the National Business Travel Association who did not work for airlines.

interviewed felt that the frequent-flier programs resulted in an average "waste" of about 8 percent of all of their travel expenditures.[22]

Frequent-flier programs put business travelers in a game situation that benefits the airlines at the expense of business travelers and their firms by encouraging noncooperative behavior. Recognizing this game, and the noncooperative incentives built into it, is important for managers who are trying to cut travel costs. And in the effort to cut these costs, managers are also in a game with the airlines, which respond to cost cutting measures with new wrinkles designed to intensify the prisoners' dilemma faced by business travelers. For example, USAir announced plans to provide a Business Select class (featuring roomier seats and better meals) for those business travelers who pay full fare for their coach tickets.[23] Of course, when all airlines have frequent-flier programs, the problems for firms may be compounded by the fact that all airlines have more "loyal" customer bases and all are less likely to cut prices (another topic to be addressed later in greater detail).

The Moral Sense

Much our analysis in the "Manager's Corner" sections of the chapters to this book will be grounded, as it has already, in the *principal/agent problem*, or the tendency of underlings to pursue their own private goals at the expense of the goals of the firm and its owners. We do that for a simple reason: We want to understand how employees *might* behave in order that managers can draw up policies and incentives that can protect the firm and its owners from agency costs.

We do not by any means wish to suggest that people are not, in the slightest degree, driven by an innate sense of duty or obligation to do that which they are supposed to do as a employee in a team or firm. On the contrary, people do seem to have a built-in tendency to cooperate -- to a degree. UCLA business professor James Q. Wilson has shown, with reference to casual observation and to a host of psychological experiments, that most people do have a "moral sense," which can show up in their willingness to forgo individual advantage (or opportunities to shirk) for the good of the group, which can be a firm.[24]

Moreover a variety of factors -- including considerations of equity and fairness -- influence people's willingness to cooperate. As organizational behaviorists have shown, "culture" has an impact on the extent of cooperation. People from "collectivistic" societies, like China, may be more inclined to cooperate than people from "individualistic" societies, like the United States.[25] Training in "group values" can affect the extent of cooperation. Experiments

---

[22] Ibid., p. 41.
[23] Ibid., p. 43.
[24] James Q. Wilson, The Moral Sense (New York: Free Press, 1993).
[25] See P. Christopher Earley, "Social Loafing and Collectivism: A Comparison of the United States and the People's Republic of China," Administrative Science Quarterly, vol. 34, n4 (Dec, 1989), pp. 565-582.

have shown that people will be more cooperative with more equal shares of whatever it is that is being divided (and women are more inclined to favor "equal shares" than men). People are willing to extend favors in cooperative ventures in the knowledge that the favor will be returned. They will work harder when they believe they are not underpaid. People are more likely to cooperate with close family members and friends than far-removed strangers, and they will be less likely to cooperate with others, whether close at hand or far removed, when the cost of cooperating is high. They work harder, in other words, when they believe they are among members of their relevant "in-group." Even training can be more effective in raising worker productivity when it is provided within in-groups, regardless of whether they come from collectivistic or individualistic societies.

Why is it that people are inclined to cooperate more or less naturally? Wilson repeats a favorite example of game theorists to explain why "cooperativeness" might be partially explained as an outcome of natural selection. Consider two people in early times, Trog and Helga, who are subject to attack by sabretooth tigers. The "game" they must play in the woods is a variant of the prisoner's dilemma game. If they both run, then the tiger will kill and eat the slowest runner. If they both stand their ground -- and cooperate in their struggle – then perhaps they can defeat the tiger. However, each has an incentive to run when the other stands his or her ground, leaving the brave soul who stands firm to be eaten.

What do people do? What *should* they do? Better yet, what do we *expect* them to do -- eventually? We suspect that different twosomes caught in the woods by sabretooth tigers over the millenniums have tried a number of strategies. However, running is, over the long run, a strategy for possible extinction, given that the tiger can pick off the runners one by one. We should not be surprised that human society has come to be dominated by people who have a "natural" tendency to cooperate or who have found ways to inculcate cooperation in their members. Moreover, parents spend a lot of family resources trying to ensure that children see the benefits of cooperation, and school teachers and coaches reinforce those values with an emphasis on the benefits of sharing and doing what one is supposed to do or has agreed to do *vis a vis* people beyond the reach of the family. Managers do much the same.

Those societies that have found ways of cooperating have prospered and survived. Those that haven't have languished or retrogressed into economic oblivion, leaving the current generation with a disproportionate representation from groups that have been cooperative. Those who didn't cooperate long ago when confronted with attacks by sabretooth tigers were eaten; those who did cooperate with greater frequency lived to propagate future generations.

What we are saying here is that human society is complex, driven by a variety of forces -- based in both psychology and economics -- that vary in intensity with respect to one another and that are at times conflicting. However, there are evolutionary reasons, if nothing else, to expect that people who cooperate will be disproportionately represented in societies that survive. Organizations can exploit -- and, given the forces of competition, must exploit -- people's limited but inherent desire or tendency to work together, to be a part of something that

is bigger and better than they are. Organizations should be expected to try to reap the synergetic consequences of their individual and collective efforts.

However, if that were the whole story -- if all that mattered were people's tendencies to cooperate -- then management would hardly be a discipline worthy of much professional reflection. There would be little or no need or role for managers, other than that of cheerleader. The problem is that firms are also beset with the very incentive problems that we have stressed. The evolutionary process is far from perfect. Moreover, as evolutionary biologist Richard Hawkins has argued, we are all beset with "selfish genes" intent on using "survival machines" (living organisms such as human beings) to increase our chances (the genes' individual chances, not so much the species' chances) of survival.[26] "Selfish genes" are willing to cooperate, if that's what is needed (or, rather, is what works); but the fundamental goal is survival. To the extent that Hawkins is right, what he might be saying is, in essence, that we have to work very hard to override basic, self-centered drives at the core of our being.

It may well be that two people can work together "naturally," fully capturing their synergetic potential. The same may be said of groups of three and four people, maybe ten or even thirty. The point that emerges from the "logic of collective action" is that as the group size -- team or firm -- gets progressively larger, the consequences of impaired incentives mount, giving rise to the growing prospects that people will shirk or in other ways take advantage of the fact that they and others cannot properly assess what they contribute to firm output.

As we have already studied, economists concerned with the economics of politics have long recognized how the "logic of choice" within groups applies to politics. The infamous "special interest" groups, which are relatively small and have long been the whipping boys of commentators, tend to have political clout that is disproportionate to their numbers. Indeed, special interest groups often get benefits from governments, with the high costs of their programs diffused over a much larger number of a more politically latent group, the general population of voters. Mancur Olson cites farmers for being the classic case of an interest group that constitutes a minor fraction (less than three percent) of the population but that has persuaded Congress to pass a variety of programs over the years that benefit farmers and their families and impose higher prices on consumers and higher taxes on taxpayers.[27]

Political economist James Buchanan points out that honor codes, which, when they work, can be valuable to all students, tend to break down as universities grow in size. For that matter, crime, which is a violation of the cooperative tendency of a community, if not a nation, tends to rise disproportionately to the population. Buchanan's explanation is that the probability

---

[26] Richard Hawkins, The Selfish Gene (New York: Oxford University Press, 1989).

[27] Mancur Olson, The Logic of Collective Action: Public Goods and the Theory of Groups (Cambridge, Mass.: Harvard University Press, 1965).

of criminals being detected, arrested, and prosecuted falls with the growth in the populations of cities.[28]

James Wilson also stresses that experimental evidence shows that people in small towns are, indeed, more helpful than people in larger cities, and the more densely packed the city population, the less helpful people will be. Presumably, people in smaller cities believe that their assistance is more detectable. People in larger cities are also less inclined to make eye contact with passersby and to walk faster, presumably to reduce their chances of being assaulted by people who are more likely to commit crimes.[29]

In his survey of the literature on the contribution of individuals to team output, Gary Miller reports that when people think that their contribution to group goals, for example, pulling on a rope, cannot be measured, then individuals will reduce their effort.[30] When members of a team pulling on a rope were blind folded and then told that others were pulling with them, the individual members exerted 90 percent of their best individual effort when one other person was supposed to be pulling. The effort fell to 85 percent when two to six other players were pulling. The shirking that occurs in large groups is now so well documented that it has a name -- "social loafing."

A central point of this discussion is not that managers can never expect workers to cooperate. We have conceded that they will – but only *to a degree, given normal circumstances*. However, there are countervailing incentive forces, which, unless attention is given to the details of firm organization, can undercut the power of people's natural tendencies to cooperate and achieve their synergetic potential.

What Firms Should Do

An important message of this chapter is that because people can't have everything they want, they will do what they can to get as much as they can. "Firms" are a means by which people can get "more" of what they want than otherwise. Firms are expensive operations, by their nature. Accordingly, people would not bother organizing themselves into "firms" if there were not gains to be had by doing so. But therein lies a fundamental dilemma for managers, how can managers ensure that the gains that could be had are actually realized and are shared in some mutually agreed upon way by all of the "stakeholders" in the firm? The problem is especially difficult when everyone associated with the firm – owners, managers, line workers, buyers, and suppliers -- probably want to take a greater share of the gains than they are getting and

---

[28]See James M. Buchanan, "Ethical Rules, Expected Values, and Large Numbers," Ethics, vol. 76 (October 1965), pp. 1-13. From the strictly economic perspective, what is truly amazing in large cities is not how many crimes are committed, but how many people respect the property and human rights of their fellow citizens, in spite of the decreased incentives to do so.

[29] Wilson, The Moral Sense, p. 49.

[30] Gary J. Miller, Managerial Dilemmas: The Political Economy of Hierarchy, (New York: Cambridge University Press, 1992), chap. 9.

contribute less in the way of work and investment than they are contributing. Managers have to find ways of overcoming the stakeholders' inclination to "give little but take a lot." One of the rolls of incentives is to overcome that inclination by tying how much people receive with what they give to the firm.

One of the more important lessons business people learn is that efficiencies can be realized from specialization and exchange. Anyone who attempted to produce even a small fraction of what he or she consumed would be a very poor person indeed.

You may recall that the late economist Leonard Reed wrote a famous article (included at the end of Chapter 1) in which he pointed out that no one person could make something even as simple as a lead pencil.[31] It takes literally thousands of people specializing in such things as the production of paint, graphite, wood products, metal, machine tools, and transportation to manufacture a pencil and make it conveniently available to consumers. No one knows enough – or can know enough – to do everything required in pencil production. Prosperity depends on our ability to become very efficient in a specialized activity and then to exchange in the market place the value we produce for a wide range of products that have been efficiently produced by other specialists. Our ability to exchange in the market place not only allows us to produce more value through specialization, it also allows us to obtain the greatest return for our specialized effort by imposing the discipline of competition on those from whom we buy.

In this chapter, we extend our discussion of how transaction costs in markets can cause firms to extend the scope and scale of their operations. We are concerned with a special form of "opportunistic behavior" relating to the use of specialized plant and equipment that can cause firms to make things themselves even though outside suppliers could produce those things more efficiently.

Make or Buy Decisions

Much the same advantage from specialization and exchange applies to firms as well as individuals. But that comment begs an important question: Exactly what should firms make inside their organizations and what should they buy from some outside vendor? Business commentators have a habit of coming up with rules that don't add very much to the answer. For example, one CEO deduced, "You should only do, in-house, what gives you a competitive advantage."[32] Okay, but why would anyone get a competitive advantage by doing anything inside, given that such a move reduces, to one degree or another, the advantage of buying from

---

[31]See Leonard Reed, "I Pencil," The Freeman, December 1958: pp. 32-37.

[32]Al Dunlap and Bob Andelman, Mean Business: How I Save Bad Companies and Make Good Companies Great (New York: Times Books, 1996), p. 55.

the cheapest outside competitor? Answers have varied over time (although the one we intend to stress relates to incentives).

At one time, the answer to the make-or-buy problem would have focused on technological considerations: Firms often produce more than one product because of what economists call "economies of scope," a situation where the skills developed in the production of one product lower the cost of producing other products.[33] But even firms with diverse product lines are actually quite specialized in that they purchase most of the inputs they use in the market rather than produce them in-house. General Motors, for example, does not produce its own steel, tires, plastic, or carpeting. Instead, it is cheaper for General Motors, and the other automobile manufacturers, to purchase these products from firms that specialize in them and to concentrate on the assembly of automobiles.[34] Neither do restaurants typically, grow their own vegetables, raise their own beef, catch their own fish, or produce their own toothpicks.

Given the advantages of specialization in productive activities and buying most of the needed inputs in the market place, a reasonable question is why firms do as much as they do? Why don't firms buy almost all the inputs they need, as they need them, from others and use them to add value in very specialized ways? Instead of having employees in the typical sense, for example, a firm could hire workers on an hourly or daily basis at a market-determined wage reflecting their alternative value at the time. Instead of owning and maintaining a fleet of trucks, a transport company could rent trucks paying only for the time they are in use. Loading and unloading the trucks could be contracted out to firms that specialize in loading and unloading trucks. The transport firm would specialize in actually transporting products. Similarly, the paper work required for such things as internal control, payroll, and taxes could be contracted out to those who specialize in providing these services.

Indeed, taken to the limit there would cease to be firms as we typically think of them. Rather there would be only individual resource owners all operating as independent contractors, with each buying (or renting) everything they need to add value in a very specialized way and then, after the value is added, selling to another individual who adds more value until a good or service is finally sold to the final consumer.

This extreme form of specialization and reliance on market exchange is clearly not what we observe in the economy. There are limits to the efficiency to be realized from further

---

[33] For example, a firm that has the equipment necessary to produce one type of electrical appliance may find that this equipment can be fully utilized if also used to produce other types of electrical appliances.

[34] Historically, automobile manufacturers did produce quite a lot of their parts in-house for reasons that will be explained later in this chapter. But the trend has been to rely more on outside suppliers, with the lowest cost manufacturers leading this trend. For example, Chrysler, the lowest-cost American producer, was producing only 30 percent of its parts in-house in the mid 1990s, versus 50 percent for Ford (the second lowest-cost American producer) and 70 percent for General Motors. Toyota produces only 25 percent of its parts in-house. See John A. Byrne, "Has Outsourcing Gone Too Far?" Business Week, April 1, 1996, p. 27.

specialization, and as a manager it is useful to understand the cause of these limits and what it implies about the advantages of producing in-house rather than buying in the market.

The problem with total reliance on the market should now be familiar: there are significant costs -- transaction costs -- associated with making market exchanges. You have to identify those who are able and willing to enter into a transaction, negotiate the specific terms of the transaction and how those terms might change under changing circumstances, draw up a contract that reflects as accurately as possible the agreed upon terms, arrange to monitor the performance of the other party to make sure the terms of the agreement are kept, and be prepared to resolve conflicts that arise between the agreement and the performance. Because of these transaction costs, it is often better for some individual or some group of individuals to directly manage the use of a variety of resources in a productive enterprise that we call a firm.

Transaction costs are lower, for example, when owners of labor become employees of the firm by entering into long-run agreements to perform tasks, that are not always spelled out clearly in advance, under the direction of managers in return for a fixed wage or salary. A market transaction is not needed every time it is desirable to alter what a worker does. Employment contracts typically allow managers wide discretionary authority to re-deploy workers as circumstances change without having to incur further transaction costs. Furthermore, with a uniform employment contract with a large number of workers, a manager can direct productive interactions between these workers that might otherwise require negotiated agreements between each pair of workers. As an example, ten workers could be hired with ten transactions, each negotiated through a relatively simple and uniform employment agreement. If those ten workers were independent contractors who had to interact with each other in ways that employees of a firm often do, they might well have to negotiate the terms of that interaction in 45 separate agreements.[35]

In general, the higher the cost of transacting through markets, the more a firm will make for itself with its own employees rather than buy from other firms. The reason restaurants don't make their own toothpicks is that the cost of transactions is extremely low in the case of toothpicks. It is hard to imagine the transaction costs of acquiring toothpicks ever getting so high that restaurants would make their own. But one might have thought the same about beef until McDonalds opened an outlet in Moscow. Because of the primitive nature of markets in Russia when McDonalds opened its first Moscow outlet (before the collapse of the Soviet Union), relying on outside suppliers for beef of a specified quality was highly risky. Because of the high transaction costs, McDonalds raised it own cattle to supply much of its beef requirements for its Moscow restaurant.

---

[35] In general N people can pair off in [(N-1)xN]/2 different ways. So ten people can pair off in [9x10]/2 = 45 different ways. The difference between the number of people (number of contracts required in an employment relationship) and the number of pairs of people (the number of contracts that could be required otherwise) increases as the number of people increases. For example, with 100 people, the number of possible pairs is 4,950. And the number of separate contracts could be larger than the number of pairs of people if they also grouped into teams with different teams having to negotiate with one another.

Negotiating an agreement between two parties can be costly, but the most costly part of a transaction often involves attempts to avoid opportunistic behavior by the parties after the agreement has been reached. Agreements commonly call for one or both parties to make investments in expensive plant and equipment that are highly specific to a particular productive activity. Once the investment is made, it has little, if any value in alternative activities. Investments in highly specific capital are often very risky, and therefore unattractive, even though the cost of the capital is less than it is worth. The problem is that once someone commits to an investment in specific capital to provide a service to another party, it is very tempting for that other party to take advantage of the investor's inflexibility by paying less than the original agreement called for.[36] There are so-called "quasi rents" that are appropriable, or that can be taken by another party through unscrupulous, opportunistic dealing.[37] The desire to avoid this risk of opportunistic behavior can be a major factor in a firm's decision to make rather than buy what it needs.

Consider an example of a pipeline to transport natural gas to an electric generating plant. Such a pipeline is very expensive to construct, but assume that it lowers the cost of producing electricity by more than enough to provide an attractive return on the investment. To be more specific, assume that the cost of constructing the pipeline is $1 billion. Assuming an interest rate of 10 percent, the annual capital cost of the pipeline is $100 million.[38] Further assume that the annual cost of maintaining and operating the pipeline is $25 million. Obviously it would not pay investors to build the pipeline for less than a $125 million annual payment, but it would be attractive to build it for any annual payment greater than that.[39] Finally, assume that if the pipeline is constructed it will lower the cost of producing electricity by $150 million dollars a year. The pipeline costs less than it saves and is clearly a good investment for the economy. But would you invest your money to build it?

---

[36] Similarly, a firm that invests in a facility that, because of its location, is dependent on a particular supplier for an important input may find that the supplier demands a higher price than agreed upon after the facility is built.

[37] For those knowledgeable in economic jargon, appropriable "quasi rents" are not the same thing as "monopoly rents" (or monopoly profits achieved by charging higher than competitive prices because of barriers to entry). Appropriable quasi rents are the differences between the purchase and subsequent selling price of an asset, when the selling price is lower than the purchase price simply because of the limited resale market for the asset. See Benjamin Klein, Robert Crawford, and Armen Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," Journal of Law and Economics (October 1978): pp. 297-326.

[38] Technically this assumes that the pipeline lasts forever. While this assumption is obviously wrong, it doesn't alter the cost figure much, if the pipeline lasts a long time. The assumption helps us simplify the example without distorting the main point.

[39] The 10 percent interest rate is assumed to be an investor's opportunity cost of capital investment. So any return greater than 10 percent is sufficient to make an investment attractive. It is assumed that the annual $25 million for maintaining and operating the pipeline includes all opportunity costs (if the payments to compensate the investor for maintenance and operation costs are made as these costs are incurred, then the costs for these items are not affected by the interest rate).

Any price between $125 and $150 million a year would be attractive to both investors in the pipeline and the electric generating plant that would use it. If, for example, the generating plant agrees to pay investors $137.5 million each year to build and operate the pipeline, both parties would realize annual profits of $12.5 million from the project. But the investors would be taking a serious risk because of the lack of flexibility after the pipeline is built. The main problem is that a pipeline is a *dedicated* investment, meaning there is a big difference in the return needed to make the pipeline worth building and the return needed to make it worth operating after it is built. While it takes at least $125 million per year to motivate building the pipeline, once it has been built it will pay to maintain and operate it for anything more than $25 million. Why? Because that is all it takes to operate the line. The pipeline investment itself is **a sunk cost**, literally and figuratively, not to be recaptured once it has been made. So after investors have made the commitment to construct the pipeline, the generating plant would be in a position to capture almost the entire value of initial pipeline investment by repudiating the original agreement and offering to pay only slightly more than $25 million per year.[40]

Of course, our example is much too extreme. The generating plant is not likely to risk its reputation by blatantly repudiating a contract. And even if it did, the pipeline investors would have legal recourse with a good chance of recovering much, if not all, of their loss. Furthermore, as the example is constructed, the generating plant has more to lose from opportunistic behavior by the pipeline owners than vice versa. If the pipeline refuses service to the plant, the cost of producing electricity increases by $150 million per year. So the pipeline owners could act opportunistically by threatening to cut off the supply of natural gas unless they receive an annual payment of almost $150 million per year.

But our main point dare not be overlooked and should be taken seriously by cost minimizing and profit maximizing business people: *Anytime a transaction requires a large investment in dedicated capital, there is the potential for costly problems in negotiating and enforcing agreements.* True, **opportunistic behavior** (actions taken as a consequence of an investment that has been made and cannot be recaptured) will seldom be as blatant as in the above example where it is clear that a lower price is a violation of the contract. But in actual contracts involving long-term capital commitments, unforeseen changes in circumstances (higher costs, interrupted supplies, stricter government regulations, etc.) can justify changes in prices, or other terms of the contract. Typically contracts will attempt to anticipate some of these changes and incorporate them into the agreed upon terms, but it is impossible to anticipate and specify

[40] Economists refer to this as capturing all the quasi rents from the investment. To elaborate on what we have already said about quasi rents, rent is any amount in excess of what it takes to motivate the supply of a good or service before any investment has been made. In the case of the pipeline, anything in addition to $125 million a year is rent. On the other hand, a quasi rent is any amount in excess of what it takes to motivate the supply of a good or service after the required investment is made. In the pipeline example, anything in excess of $25 million a year is quasi rent. So once the investor has committed to the pipeline, any offer over $25 million a year will motivate the supply of pipeline service and allow the generating plant to capture almost all of the quasi rent.

appropriate responses to all possible changes in relevant conditions. Therefore, there will usually be ambiguities in long-term contractual arrangements that open the door for opportunistic behavior of the type just discussed, and that can be resolved only through protracted and expensive legal action.

So committing to investments in dedicated capital carries great risk of opportunistic behavior without some assurance that such behavior will not pay. One way to obtain this assurance is for the investment to be made by the same firm that will be using the output it produces. Alternatively, the firm that makes the investment in the specific capital can merge with the firm that depends on the output from that investment.

The early history of the automobile industry provides an example of a merger between two companies that can be explained by the advantages of producing rather than buying when dedicated capital investment is involved.[41] In 1919, General Motors entered into a long-term contract with Fisher Body for the purchase of closed metal car bodies. This contract required that Fisher Body invest in expensive stamping machines and dies specifically designed to produce the bodies demanded by GM. This put Fisher Body in a vulnerable position, given that once the investment was made GM could have threatened to buy from someone else unless Fisher Body reduced prices substantially. This problem was anticipated, which explains why the contract required that GM buy all of the closed metal bodies from Fisher and specified the price as equal to Fisher's variable cost plus 17.6 percent.

However, while these contractual terms protected Fisher against opportunistic behavior on the part of GM, they created an unanticipated opportunity for Fisher to take advantage of GM. The demand for closed metal bodies increased rapidly during the early 1920s (in part because of increased auto sales, but also from a dramatic shift from open wooden bodies to closed metal bodies). The increased production lowered Fisher's production costs, and indeed made it possible for Fisher to lower its costs significantly more than it did. Evidence suggests that Fisher took advantage of the 17.6 percent "price add-on" by keeping its variable costs (particularly labor costs), and therefore the price charged GM, higher than necessary.

General Motors was aware of this "over charge" and requested that Fisher build a new auto body plant next to GM's assembly plant. This would have eliminated the costs of transporting the auto bodies (a variable cost that came with the 17.6 percent add-on) and reduced GM's price. Fisher refused to make the move, however, possibly because of concerns that such a dedicated investment to GM requirements would be exploited by GM. As a result of the potential haggling, threats and counter-threats, GM bought Fisher Body in 1926 and the two companies merged. GM could buy Fisher simply because their tenuous dealings, with accompanying transaction costs, were depressing both companies' market value. GM could pay a premium for Fisher simply because of the anticipated transaction cost savings.

---

[41] The following discussion of the relationship between General Motors and Fisher Body is taken from Klein, Crawford, and Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," Journal of Law and Economics (October 1978): pp. 308-310.

In an ideal world without transaction costs, General Motors would have bought auto bodies from specialists subject to the constant discipline of market competition. In the real world of transaction costs, GM made the auto bodies itself.

The construction of electric generating plants next to coalmines provides another example of the potential benefits to a firm for producing an input rather than buying it when highly specific capital is involved. There is an obvious advantage in "mine-mouth" arrangements from reducing the cost of transporting coal to the generating plant. But if the mine and the generating plant are separately owned, the potential for opportunistic behavior exists after the costly investments are made. The mine owner, for example, could take advantage of the fact that the generating plant is far removed from a rail line connecting it to other coal supplies by increasing the price of coal. To avoid such risks, common ownership of both the mine and the generating plant is much more likely in the case of "mine-mouth" generating plants than in the case of generating plants that can rely on alternative sources of coal. And, when ownership is separate in a "mine-mouth" arrangement, the terms of exchange between the generating plant and mine are typically spelled out in very detailed and long-term contracts that cover a wide range of future contingencies.[42]

There are other ways a firm can benefit from the advantages of buying an input rather than producing it while reducing the risks of being "held-up" by a supplier who uses specialized equipment to produce a crucial input. It can make sense for the firm to buy the specialized equipment and then rent it to the supplier. If the supplier attempts to take advantage of the crucial nature of the input, the firm can move the specialized equipment to another supplier rather than be forced to pay a higher than expected price for the input. This is exactly the arrangement that automobile companies have with some of their suppliers. Ford, for example, buys components from many small and specialized companies, but commonly owns the specialized equipment needed and rents it to the contracting firms.[43]

Firms are also aware that those who supply them with services are reluctant to commit themselves to costly capital investments that, once made, leave them vulnerable to **hold-up** (demands that the terms and conditions of the relationship be changed after an investment that cannot be recaptured has been made). In such case the firm that provides the capital equipment

---

[42] For a detailed discussion of the mine-mouth arrangements, see Paul Joskow, "Vertical Integration and Long-Term Contracts: The Case of Coal-Burning Electric Generating Plants," Journal of Law, Economics, and Organization (Spring 1985): pp. 33-80.

[43] The Ford example is discussed on pages 245-46 of Robert Cooter and Thomas Ulen, Law and Economics (Glenview, Illinois: Scott, Foresman and Company, 1988). Also, Alex Taylor III, op cit., discusses this strategy by automobile companies as a way of reducing the number of suppliers they depend on (therefore reducing transaction cost) without increasing their vulnerability to hold-up. On p. 54 he states, "Even now some manufacturers pay for the suppliers' equipment so if production falters, they can yank out the machinery and install it in someone else's factory." These arrangements also have advantages from the small contracting companies' perspective, since they provide a signal to the auto companies that the contractors will play straight with them. The advantage of a business being able to commit itself to honest dealing is discussed later in the book.

and rents it to the supplier can benefit from the fact that less threatened suppliers will charge lower prices. This consideration may also be a motivation for auto manufacturers to own the equipment that some of their suppliers use. It also provides a very good incentive-based explanation, and justification, for a business arrangement that has been widely criticized.

An arrangement that reduced the threat of opportunistic behavior on the part of firms against workers was the much-criticized "company town." In the past it was common for companies (typically mining companies) to set up operations in, what were at the time, very remote locations. In the company towns, the company owned the stores where employees shopped and the houses where they lived. The popular view of these company stores and houses is that they allowed the companies to exploit their workers with outrageous prices and rents, often charging them more for basic necessities than they earned from backbreaking work in the mines. The late Tennessee Ernie Ford captured this popular view in his famous song "Sixteen Tons."[44]

Without denying that the lives of nineteenth-century miners were tough, company stores and houses can be seen as a way for the companies to reduce (but not totally eliminate) their ability to exploit their workers by behaving opportunistically. Certainly workers would be reluctant to purchase a house in a remote location with only one employer. The worker who committed to such an investment would be far more vulnerable to opportunistic wage reductions by the employer than would the worker who rented company housing. Similarly, few merchants would be willing to establish a store in such a location, knowing that once the investment was made they would be vulnerable to opportunistic demands for price reductions that just covered their variable costs, leaving no return on their capital cost. Again, in an ideal world without transaction costs – and without opportunistic behavior -- mining companies would have specialized in extracting ore and would have let suppliers of labor buy their housing and other provisions through other specialists. But in the real world of transaction costs, it was better for mining companies to also provide basic services for their employees. This is not to say that there was no exploitation. But the exploitation was surely less under the company town arrangement than if, for example, workers had bought their own houses.[45]

The threat to one party of a transaction from opportunistic behavior on the part of the other party explains other business and social practices. Consider the fact that despite valiant efforts, the vast majority of farm workers have never been able to effectively unionize in the United States. No doubt many reasons explain this failure, but one reason is that a union of farm workers would be in a position to harm farmers through opportunistic behavior. A crop is a highly specialized and, before harvested, immobile investment, and one whose value is easy to

---

[44] The lyrics of which went, "Sixteen tons and what do you get? Another day older and deeper in debt. Saint Peter don't you call me cause I can't go. I owe my soul to the company store."

[45] For a relevant discussion of company towns set up by coal mining firms, see Price V. Fishback, Soft Coal, Hard Choices: The Economic Welfare of Bituminous Coal Miners, 1890-1930 (New York: Oxford University Press, 1992); especially chapters 8 and 9.

expropriate at harvest time. In most cases, if a crop is not harvested within a short window of opportunity, its value perishes. Therefore, a labor union could use its control over supply of farm workers to capture most of a crop's value in higher wages by threatening to strike right before the harvest. While this threat would not necessarily be carried out in every case, it is too serious for those who have made large commitments of capital to agricultural crops to ignore. Not surprisingly, farm owners have strongly resisted the unionization of farm workers.

The threat of opportunistic behavior is surely an important consideration in another important exchange relationship, that of marriage. Although there clearly are exceptions, rich people seldom marry poor people. The story of the wealthy prince marrying poor, but beautiful, Cinderella, is, after all, a fairy tale. Rich people generally marry other rich people. As with all activities, there are many explanations for marital sorting, including the obvious fact that the rich tend to hang around others who are rich. But an important explanation is that marriage is effectively a specialized investment that, once made, commits and creates value not easily shifted to another enterprise, or object of affection. The rich person who marries a poor person is making an investment that is subject to hold-up. This is a hold-up possibility that is not ignored, as evidenced by the fact that pre-nuptial agreements are common in the case of large wealth differences between the two parties to a marriage. But because of the difficulty of anticipating all possible contingencies relevant to distributing wealth upon the termination of a marriage, such agreements still leave lots of room for opportunistic behavior. Marriage between people of roughly equal wealth reduces, though hardly eliminates, the ability of one party to capture most of the value committed by the other party.

A good general rule for a manager is to buy the productive inputs the firm needs rather than make them. When inputs are produced in-house, some of the efficiency advantages of specialization provided through market exchange are lost. But as with most general rules, there are lots of exceptions to that of buying rather than making. In many cases the loss from making rather than buying will be more than offset by the savings in transaction costs. Typically, firms should favor making those things that require capital that will be used for specific purposes and, therefore, will not have a ready resale market.

The Decision to Franchise

The decision a firm faces over whether to expand through additional outlets that are owned by the firm or that are franchised to outside investors has many of the features of decisions to make or buy inputs. Franchising is simply a type of firm expansion – with special contractual features and with all the attendant problems. Franchise contracts between the "franchiser" (franchise seller) and the "franchisee" (franchise buyer) typically have several key features:

- The franchisee generally makes some up-front payment, plus some royalty that is a percentage of monthly sales, for the right to use a brand name and/or trademark -- for example, the name "McDonalds" along with the "golden arches."

- The franchisee also agrees to conduct business along the lines specified by the franchise, including the nature and quality of the good or service, operating hours, sources of purchases of key resources in the production process, and the prices that will be charged.
- The franchiser, on the other hand, agrees to provide managerial advice and to undertake advertising, to provide training, and to ensure that quality standards are maintained across all franchisees.
- The franchiser typically retains the right to terminate a franchise agreement for specified reasons, if not at will.

The own-or-franchise decision is similar to the make-or-buy decision because both types of decisions involve problems of monitoring, risk sharing, and opportunistic behavior. At one time, scholars believed that firms expanded by way of franchising only as a means of raising additional capital through tapping the franchisee's credit worthiness. If the firm owned the additional outlet, it would have to bring in more investors or lenders at higher capital costs. Supposedly, franchisees could raise the money more cheaply than the franchiser.[46]

However, Emory University economist Paul Rubin has argued with force that franchising, per se, doesn't, and can't, reduce the overall cost of capital – at least not as directly as previously argued.[47] A firm in the restaurant business, for example, can only contemplate expanding through franchising if it has a successful anchor store. It can establish another outlet through the sale of its own securities, equities or bonds, in which case the investors will have an interest in both the successful anchor restaurant and the new one. That investment in a combination of the proven and new restaurant is likely to be less risky than any single investment in just the new restaurant, which, because it has the same menu as the anchor restaurant, has a good chance of success, but is still unproved. Hence, the cost of capital for the franchisee, everything else held constant, is likely to be higher than for the central restaurant firm.

Why franchise ever? Rubin argues that in business there are unavoidable **agency costs**, or costs associated with the fact that the owners (or **principals**) of a firm must hire managers and workers (**agents**) who have discretion in the use of firm resources but who do not necessarily have the right incentives to use the firm's resources in the most effective manner to pursue the owners' goals, as opposed to the private goals of the managers and workers. Rubin believes the reason for franchising is that the **agency cost** is lowered (but not totally eliminated) by expanding through franchising. The manager of the company-owned restaurant will likely be paid a salary plus some commission on (or bonus related to) the amount of business. The manager's incentive will be weakly related to the interests of the owners. Hence,

---

[46]This argument is evident in Donald N. Thompson, Franchise Operations and Antitrust (Lexington, Mass.: D.C. Heath, 1971).

[47]Paul H. Rubin, "The Theory of the Firm and the Structure of the Franchise Contract," Journal of Law and Economics, vol. 21 (1978), pp. 223-233.

the manager will have to be closely monitored. The franchisee, on the other hand, becomes the residual claimant on the new restaurant business and, accordingly, has a stronger incentive to reduce shirking and other forms of opportunistic behavior by the employees.

We note above that **monitoring costs** (or the costs associated with keeping track of manager and worker performance) are *not eliminated* through franchising. This is the case because the franchisees have some reason to shirk (albeit that the incentive to shirk is impaired by the franchise agreement that leaves the franchisee an important residual claimant). Customers often go to franchised outlets because they have high confidence in the nature and quality of the goods and services offered. McDonalds customers know that they may not get the best burger in town when they go to a McDonalds, but they do have strong expectations on the size and taste of the burgers and the cleanliness of the restaurant. McDonalds has a strong incentive to build and maintain a desired reputation for its stores, and therein lies the monitoring catch. Each franchisee, especially those that have limited repeat business, can "cheat" (or free ride on McDonalds overall reputation) by cutting the size of the burgers or letting their restaurants deteriorate. The cost savings for the individual cheating store can translate into a reduced demand for other McDonalds restaurants. This is a prisoner's dilemma in which all stores can be worse off if noncooperative behavior becomes a widespread problem. So, McDonalds must set (and has strong incentives to do so) production and cleanliness standards and then back up the standards with inspections and fines, if not outright termination of the franchise contract.

McDonalds (and any other franchiser) also controls quality by requiring the individual restaurants to buy their ingredients -- for example, burger patties and buns -- from McDonalds itself or from approved suppliers. McDonalds has good reason to want its franchisees to buy the ingredients from McDonalds, not because (contrary to legal opinion) it gives McDonalds some sort of monopoly control, but because McDonalds has a problem in monitoring outside suppliers.[48] Outside suppliers have an incentive to shirk on the quality standards with the consent of the franchisees that, individually, have an interest in cutting their individual costs. Moreover, by selling key ingredients, the franchiser has an indirect way of determining if its royalties are being accurately computed. So-called "tie-in sales" are simply a means of reducing monitoring costs. Of course, the franchises also have an interest in their franchiser having the lowest possible monitoring cost: it minimizes the chances of free riding by the franchisees and maintains the value of the franchise. Similarly, a franchiser like McDonalds (as do the franchisees) has an interest in holding all franchisees to uniform prices that are higher than individual McDonalds might want to choose. By maintaining uniform retail prices, McDonalds encourages its franchisees to incur the costs that must be incurred to maintain desired quality standards.

---

[48]Rubin, "The Theory of the Firm and the Structure of the Franchise Contract," p. 254. For a review of legal opinion on the so-called "tie-in sales" of franchise relationships, see Benjamin Klein and Lester F. Saft, "The Law and Economics of Franchise Tying Contracts," Journal of Law and Economics, vol. 28 (May 1985), pp. 345-361.

The chances for opportunistic behavior can be lowered through franchising, but hardly eliminated.[49] If the franchisee buys the rights to the franchise and then invests in the store that has limited resale value, the franchiser can appropriate the rents simply by demanding higher franchise payments or failing to enforce production and quality standards with the franchisees, increasing the take of the franchiser but curbing the resale value of the franchise. On the other hand, if the franchisee pays for the building that has a limited resale value, the franchisee can, after the fact, demand lower franchise fees and special treatment (to the extent the franchiser must incur a cost in locating another franchisee).

These points help explain up-front payment and royalty provisions in franchise contracts. The value of the franchise to the franchisee – and what the franchisee will pay, at a maximum, for the franchise – is equal to the present value of the difference between two income streams, the income that could be earned with and without the franchise. The greater the difference, the greater the up-front payment the franchisee is willing to make. However, the franchisee is not likely to want to pay the full difference up-front. This is because the franchiser would then have little incentive to live up to the contract (to maintain the flow of business and to police all franchisees). The franchiser could run off with all the gains and no costs. As a consequence, both the franchiser and franchisee will likely agree to an up-front payment that is less than the difference in the two income streams identified above and to add a royalty payment. The royalty payment is something the franchisee, not just the franchiser, will want to include in the contract simply because the franchiser will then have a stake in maintaining the franchisee's business. A combination of some up-front payment and royalty is likely to maximize the gains to both franchisee and franchiser.

Franchising also has risk problems no matter how carefully the contract may be drawn. Typically, franchisees invest heavily in their franchise, which means the franchisee has a risky investment portfolio because it is not highly diversified. This can mean that the franchisee will be reluctant to engage in additional capital investment that could be viewed as risky only because of the lack of spread of the investment. As a consequence, franchisers will tend to favor franchisees that own multiple outlets. A franchisee with multiple outlets can spread the risk of its investments and can more likely internalize the benefits of its investments in maintaining store quality (customers are more likely to patronize, or fail to do so, at another of the owner's outlets).

Obviously, both ownership and franchise methods of expansion have costs and benefits for investors. We can't here settle the issue of how a firm like McDonalds should expand, by ownership of additional outlets or by franchising them. All we can do is point out that franchising should not be as important when markets are "local." It should not, therefore, be a surprise that franchising grew rapidly in the 1950s with the spread of television that greatly expanded the market potential for many goods and services and when transportation costs

[49]See James A. Brickley and Frederick H. Dark, "The Choice of Organizational Forms: The Case of Franchising," Journal of Financial Economics, vol. 18 (1987), pp. 401-420.

began declining rapidly, which allowed people to move among local markets.[50] *Franchising will tend to be favored when there is a low investment risk for the franchisee and when there are few incentives for free riding by both franchisee and franchisers*. We should expect that franchises should be favored the greater the monitoring costs (implying the farther the store location is from the franchiser, the more likely the expansion will be through franchising, a conclusion that has been supported by empirical studies[51]). Also, we would expect stores at locations with relatively few repeat customers to be company owned. A better way of putting that point is the fewer the repeat customers in a given location, the greater the store will be company owned. When a store has few repeat customers, the incentive to cheat is strong, which means that the franchiser will have to maintain close monitoring to suppress the incentive for the franchisee to cheat or free ride – which implies there may be fewer cost advantages to franchising the location.[52] If monitoring costs go down, we should expect firms to increase their ownership of their outlets.

Much of what we have written in this chapter is based on the presumption that people will behave opportunistically. We see the presumption as well grounded, given the extent to which people do behave that way in their daily dealings (and most managers have no trouble identifying instances of opportunistic behavior in workers, suppliers, and investors). We may, however, have given the impression that we believe that *all* people are *always* willing to behave opportunistically, which is simply contradicted by everyday experience. The business world is full of saints and sinners, and most people are some combination of both. We simply base our discussion here and in later chapters on a presumption that people will behave opportunistically not because such an assumption is fully descriptive of everyone in business, but because that is the threat managers want to protect themselves against. Business people don't have to worry about the Mother Teresa's of the world. They do have to worry about less-than-perfect people. (And they do have to worry about people who pretend to be like Mother Teresa before any deal is consummated.) They need to understand the consequences of opportunistic behavior in order that they can appropriately structure contracts and embedded incentives.

---

[50]G. Frank Mathewson and Ralph A. Winter, "The Economics of Franchise Contracts," Journal of Law and Economics, vol. 28 (October 1985), p. 504.

[51]Brickley and Dark, "The Choice of Organizational Forms: The Case of Franchising," pp. 411-416.

[52] Unfortunately, the only available study on the relationship between the extent of repeat business and the likelihood of franchising (Brickley and Dark, "The Choice of Organizational Forms: The Case of Franchising,") does not confirm the theory. These researchers investigated how the location of outlets near freeways affected the likelihood that they would be franchised. They assumed that locations near freeways would have limited repeat business. Hence, they expected that locations near freeways would tend to be company owned, but they found the exact opposite: outlets near freeways tended to be franchised. The inconsistency between the findings and the prediction could be explained by the fact that the theory is missing something. However, it could also be, as the researchers speculate, that the problem is their measure of repeat business; locations near freeways may not be a good measure of repeat business. Such locations might get more repeat business than was assumed when it was selected as a proxy.

Here, we have shown how opportunistic behavior can arise in the most basic of management decisions, whether to "make or buy." An important task of a good manager is being constantly attentive to the trade-off between the advantages of buying and those of making, and one of the major worries is the extent of opportunistic behavior in that decision. In assessing this trade-off managers need to be aware that the decision is dependent upon the nature of what is to be bought or produced and that bureaucratic tendencies within a firm can distort decisions in favor of producing in-house even though buying would be more efficient. The firm that loses sight of this tendency may soon be out-competed by smaller firms that rely less on internal allocation and more on specialization and market transactions to produce at lower cost.

This suggests that the size and specialization of firms will change over time in response to technological advances that alter the relative costs of market transactions and the costs (as well as the efficiency) of managerial control. In other chapters we discuss the effects that improvements in communication, transportation, and management information systems are having on the size and focus of firms. The trend for firms to downsize and to refocus on their "core competencies" can be explained, at least in part, by the reduced cost of smaller, more specialized firms dealing with each other through market exchange in collaborative productive efforts. But no matter how specialized firms become, resources will continue to be allocated differently within firms than they are across markets. The reason firms will continue to exist is that over some range of productive activity, it is more efficient for resources to be directed by managerial control than by market exchange.[53]

MANAGER'S CORNER: **Fringes, Incentives, and Profits**

Varying the form of pay is one important way firms seek to motivate workers – and overcome the prisoners' dilemma/principal-agency problems that have been at the heart of this chapter. And worker pay can take many forms, from cold cash to an assortment of fringe benefits. However, it needs to be noted that workers tend to think and talk about their fringe benefits in remarkably different terms than they do about their wages. Workers who profess that they "earn" their wages will describe their fringes with reference to what their employers "give" them. "Gee, our bosses *give* us three weeks of vacation, thirty minutes of coffee breaks a day, the right to flexible schedules, and discounts on purchases of company goods. They also provide us with medical and dental insurance and cover 80 percent of the cost. Would you believe we only have to pay 20 percent!"

[53] It should be pointed out that even when managers within the firm control resources, this control couldn't be exercised independently of market forces, at least not for long. Unless the firm is using its productive resources to produce goods and services that pass the market test, it will soon be forced through bankruptcy and have to relinquish those resources to more efficient firms.

Wages are the result of hard work, but fringes, it seems, are a matter of employer generosity. Fringes are assumed to come from a substantially different source, such as out of the pockets of the stockholders, than wages, which come out of the revenues workers add to the bottom line.

Employers use some of the same language, and their answers to any question of why fringes are provided are typically equally misleading, though probably more gratuitous. The main difference is that employers inevitably talk in terms of the cost of their fringes. "Would you believe that the cost of health insurance to our firm is $4,486 *per employee*? That means that we spend millions, if not tens of millions, each year on all of our employees' health insurance. Our total fringe-benefit package costs us an amount equal to 36.4 percent of our total wage bill!" The point that is intended, though often left unstated is "Aren't we nice?"

If either the workers or the employers who make such comments are in fact telling the truth, then the company should be a prime candidate for a hostile takeover. Someone -- a more pragmatic and resourceful businessperson -- should buy the owners out, and the workers should *want* that someone to buy the company because they could then share in the gains to be had from the improved efficiency of the company.

Our arguments here will be a challenge to many readers since it will develop a radically different way of thinking about fringe benefits. It will require readers to set aside any preconceived view that fringes are a gift or that fringes are either provided or they are not. The approached used here employs what we call *marginal* analysis, or the evaluation of fringes in terms of their *marginal* cost and *marginal* value. It is grounded in the principle that profits can be increased so long as the marginal value of doing anything in business is greater than the marginal cost.

This principle implies that a firm should extend its output for as long as the marginal value of doing so (in terms of additional revenue) exceeds the marginal cost of each successive extension. It should do the same with a fringe: provide it so long as it "pays," meaning so long as the marginal cost of the fringe is less than its marginal value (in terms of wages workers are willing to forgo and greater production) for the firm. This way of looking at firm decision-making means that changes in the cost of fringes can have predictable consequences. An increase in the cost of any fringe can give rise to a cut in the amount of the fringe that is provided. An increase in the value of the fringe to workers can lead to more of the fringe being provided.

Workers As Profit Centers

We don't want to be overly crass in our view of business (although that may appear to be our intention from the words we have to use within the limited space we have to develop our arguments). We only want to be realistic when we surmise that from our economic perspective (the one that is likely to dominate in competitive business environments), the overwhelming

majority of firms that provide their workers with fringes do so for the very same reason that they hire their workers in the first place: To add more to their profits than they could if they did something else. *Like it or not, most firms are in the business of making money off their employees -- in all kinds of ways.*

The reason many firms don't provide their workers with fringe benefits -- with health insurance being the most common missing fringe in small businesses especially -- is that they can't make any money by doing so. The critical difference between those employers who do provide fringes and those who don't is not likely to have anything to do with how nice each group wants to be to its employees. We suspect that both groups are equally nice, or equally crass. There is really no reason to believe that people who do not provide some form of fringes (or provide less of some form) are, on average, any more derelict in their duty to serve mankind than are the people who do.

When making decisions on fringe benefits employers face two unavoidable *economic* catches: First, fringes are costly, and some fringes, like health and dental insurance, are extraordinarily costly. Second, there are limits to the value workers place on fringes. The reason is simply that workers value a lot of things, and what they *buy*, directly from vendors or indirectly via their employers, is largely dependent on who is the lowest cost provider.

Yes, workers *buy* fringe benefits from employers. They do so when the value the workers place on the fringes exceeds the cost of the fringes to the firms. When that condition holds, firms can make money by, effectively, "selling" fringes -- for example, health insurance -- to their workers. How? Most firms don't send sales people around the office and plant selling health insurance or weeks of vacation to their employees like they sell fruit in the company cafeteria, but they nevertheless make the sales. They do it somewhat on the sly, indirectly, by offering the fringes and letting their particular labor market conditions adjust. If workers truly value a particular fringe, then the firms that provide the fringe will see an increase in the supply of labor available to them. They will be able to hire more workers at a lower wage and/or be able to increase the "quality" (productivity) of the workers that they do hire.

Firms are paid for the cost of providing fringe benefits primarily in two ways: One, their real wage bill goes down with the increased competition for the available jobs that results from the greater number of job seekers (who are attracted by the fringe). This reflects the willingness of workers to *pay* employers for the fringe benefits. Two, employers gain by being more discriminatory in whom they hire, employing more productive workers for the wages paid and increasing sales.

No matter what happens in particular markets, we know several things about the pattern that will emerge in the fringe-benefit market:

- Many firms (but not all) can make money by "selling" fringes to their workers.
- Firms won't provide the fringes if the combined gains from lower wages and better workers are not greater than the cost of the fringes.

- Workers, who may suffer a decline in their wages because of their fringes, will still be better off because of the fringes that they buy. Otherwise, the fringes would not be made available by the firm or the number of job seekers would not increase, and the firms could not justify providing the fringe.
- If providing a given fringe is profitable for firms, there will be competitive pressures to provide it. Otherwise, firms that do not provide the fringe will have a higher cost structure (because their total wage bill will be higher by more than the cost of the fringe) and will be in a less competitive position.

To see these points with greater clarity, we must look to a graph, albeit a simple one, using only the supply and demand curves with which you must now be familiar. We have drawn in Figure 6.2 normal labor supply and demand curves. The downward sloping labor demand curve, D_1, shows that more workers will be demanded by firms at lower wage rates than higher wage rates and reflects the circumstance in which no fringe benefit is provided. The upward sloping curve, S_1, shows that more workers will come on the markets at higher wage rates than at lower ones and reflects an initial circumstance in which a given fringe benefit (such as health insurance) is not provided. These embedded assumptions regarding the slopes of the curves are totally reasonable and widely accepted as reflecting market conditions. At any rate, without the fringe the workers will receive a wage rate of W_1, where the market clears.

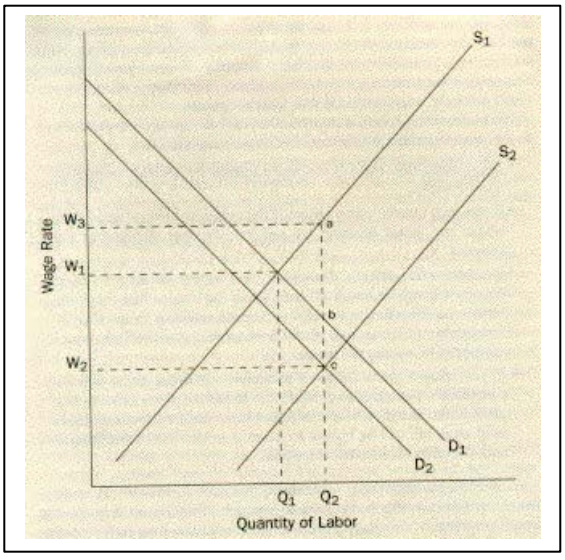


Figure 6.2 Fringes and the Labor Market

If fringes are more valuable to workers and they impose a cost on the employers, the supply of labor will increase from $S1$ to $S2$ while the demand curve falls from D_1 to D_2. The wage rate falls from W1 to W_2, but the workers get fringes that have a value of ac, which means that their overall payment goes up from W_1 to W_3.

Consider the simplest of cases, the one in which the firm's cost in providing a fringe benefit is a uniform amount for each worker and in which the provision of the fringe has no impact on worker productivity, but increases the value of work and increases the supply of

workers. The demand curve in Figure 6.2 drops down vertically by the per-worker cost of the fringe, from D_1 to D_2. This happens because the firms are simply not willing to pay as high a wage to their workers if they have to cover the cost of the fringe. On the other hand, the supply of workers shifts outward, from S_1 to S_2, because work is now more attractive because of the fringe, leading to more workers applying for jobs. Workers are willing to work for a lower money wage *when the fringe is provided* (and, again, for simplicity we assume that each worker values the fringe by the same amount). The vertical difference between S_1 and S_2 represents how much each worker values the fringe and is willing to give up in their wage rate for the fringe; this vertical difference is a money measure of the value of the fringe to workers.

What happens, given these shifts in supply and demand? As can be seen in the figure, the market-clearing wage falls from W_1 to W_2. Are workers and firms better off? Well, a close examination of the figure reveals that more workers are employed (Q_2 instead of $Q1$), which suggests that something good must have happened. Otherwise, we must wonder why firms would want to hire more workers and more workers would be willing to be employed. It just doesn't make much sense to argue that firms and/or workers aren't better off when both sides agree to more work (and when the fringe is provided voluntarily).

Notice that the total cost of the fringe, the vertical distance between the two demand curves, or *bc*, is less than the reductions in the wage, *W*1-*W*2, from which we can draw two implications: First, the firm is clearly making money off its original employees ($W_2 + bc$ is less than W_1). Second, the firm's total cost per worker ($W_2 + bc$) falls, which explains why they are willing to expand their hires.

Notice also that while the workers accept a lower wage rate, W_2 instead of W_1, they gain the value of the fringe, which in the graph is the vertical distance *ac*. The sum of the new lower wage, W_1, plus the value of the fringe, *ac*, is W3, which is higher than the wage without the fringe ($W_2 + ac = W_3 > W_1$). *Ergo*, both sides gain.

How much of the fringe benefit should be provided? It would be nice if we could tell each person reading this book exactly what to do. It would be silly to try, given the variation of business and market circumstances. What we can do is look to rules that are generally applicable. The rule the firms should follow is no different than the rule they should follow in any other productive market circumstance: *Firms should continue to expand the fringe so long as the added cost from the fringe is less than the reduction in their wage bills, which can be no greater than the workers' evaluation of the fringe.*

For example, the number of days of paid vacation should be extended so long as the value workers place on additional vacation days is greater than the marginal cost to the employer of providing the additional day. Given that workers' evaluation of each additional day will fall (at least after some number of days) and the cost of the additional day will rise, after some number of days off, a point will be reached beyond which equality between the additional cost of the next vacation day will exceed its marginal value (or the possible reduction in the

wage bill). At that point, employers have maximized their profit from "selling" the fringe to their workers.

Of course, tax rules will affect the exact amount of the fringe, as well as the combination. Certainly, if fringe benefits -- for example, health insurance -- are not subject to taxation, then employers should, naturally, provide more of them than otherwise, simply because part of the cost of the benefit is covered by a reduction in worker taxes. The result might be that workers actually get more of the benefit than they would buy, *if they were covering all of the cost themselves*. Still, the employers must provide the benefit; otherwise, they will not keep their compensation costs competitive with that of rival employers.

Optimum Fringes

We expect employers and workers to treat fringes like they do everything else, seeking some *optimum* combination of fringes and money wages. Again, this means that employers and workers should be expected to weigh off their additional (or marginal) value against their additional (or marginal) cost. An employer will add to a fringe like health insurance as long at the marginal value (measured in money wage concessions or increased production from workers) is greater than the marginal cost of the added fringe. Similarly, workers will "buy" more of any fringe from their employer so long as its marginal value (in terms of improved health or reduction in the cost of private purchase) is greater than its marginal cost (wage concessions).

While we can't give specifics, we do know that managers are well advised to search earnestly for the "optimum" combination (which means some experimentation would likely be in order) even though the process of finding the optimum is beset with imprecisions. The firms that come closest to the optimum will be the ones that can make the most money from their employees. They will also be the ones that provide their employees the most valuable compensation for the money spent -- and so will have the lowest cost structure and be the most competitive. By trying to make as much money as possible from their employees, firms not only stay more competitive, they also benefit their workers as well.

So far, we have considered only fringes in which the added cost of the fringe to the firm is less than the value of the fringe to the workers. What if that were not the case? Returning attention to Figure 6.2, suppose that the cost of the fringe to firms were greater than the value of the fringe to workers (in the graph, the distance *bc* is greater than the distance *ac*), what would happen? The straight answer: Nothing. The fringe would not be provided. The reason is obvious: Both sides, workers and owners, would lose. The resulting drop in the wage would be less than the cost of the fringe to the employers, and the resulting drop in the wage would be greater than the value of the fringe to the workers. (To see this point, just try drawing a graph with the vertical drop in the demand greater than the outward shift of the supply.) *Such a fringe would not -- and should not -- be provided simply because it is a loser to both sides.*

Firms that persisted in providing such a fringe would have difficulty competing, simply because their cost structure would be higher than other producers. Such firms would be subject to takeovers. The takeover would very likely be friendly because those bidding for the firm in the takeover would be able to pay a higher price for the stock than the going market price, which would be depressed by the fact that one or more fringes provided to workers was not profitable. Those involved in the takeover could, after acquiring control, eliminate the excessively costly fringe(s) (or reduce it to profitable levels), enhance the firm's profitability and competitive position, and then sell the firm's stock at a price higher than the purchase price.[54]

The workers would support such a takeover -- and might be the ones managing the takeover -- because they could see a couple of advantages: They could have a fringe eliminated that is not worth to them the cost that they would have to pay in terms of lower wages. They could also gain some employment security, given the improved competitive position of their firm. The workers might even take the firm over for the same reason anyone else might do so: They could improve the firm's profitability and stock price.

Fringes Provided by Large and Small Firms

We can now understand why it is that so many large firms provide their employees with health insurance and so many small firms do not. At the most general level, it simply pays large firms to provide the insurance, while it doesn't pay for the small firms to do so. Large firms can sell a large number of health insurance policies, achieving economies associated with scale and of spreading the risks. That's a widely recognized answer.

At another level, the answer is more complicated and obscure. "Small" and "large" firms do not generally hire from the same labor markets. Small firms tend to provide lower paying jobs. The workers in lower paying jobs within small firms simply don't have the means to buy a lot of things that workers in larger firms have, and one of the things workers in small firms don't seem to buy in great quantities is insurance. Given their limited income, workers simply don't think that insurance is a good deal, and they would prefer to buy other things with higher monetary compensation. One of the reasons low-income workers may gravitate to small firms is that they shy away from large firms where they would have to give up wages to buy the insurance, because of company policies that apply to all workers.

Of course, the analysis gets even trickier when it is realized that lower income workers, many of whom work for small firms, tend to be younger workers -- who also tend to be healthier and in need of a different combination of fringes than older workers. The young can appreciate that the price they would have to pay for health insurance through their firms is

[54] Engaging in a takeover can be very expensive, and we recognize that a firm is not likely to be taken over because of the failure of the firm to provide one efficiency-enhancing fringe benefit. But when enough of these types of mistakes are made, the inefficiency mounts, increasing the chance that the firm will be a takeover target.

inflated by a number of factors related to supply and demand. First, the price of health insurance has been inflated by a host of cost factors, not the least of which is the increased liability doctors face for virtually anything that goes wrong with patients when they are under the doctors' care. The radical application of expensive medical technologies to care for older dying patients has also jacked up the cost of insurance and care for the young.

Second, older workers, many of whom are in large firms and tend to have a strong demand for health insurance, have increased the demand for insurance (and health care). The exemption of health insurance from taxable income (which helps higher income workers more than lower income workers) has also artificially inflated the demand for health insurance (and health care). The net result of the cost and demand effects has been to increase health insurance costs, making the insurance an unattractive deal for many young and low-income workers, many of whom work for small firms.

We know the objections to our line of analysis. Critics might say that we have overlooked the human factor. Fringe benefits are important to workers, and they should have fringe benefits even when they aren't profitable. We see a couple of problems with that claim. If the fringe were as important as claimed, then surely workers would be willing to give up a lot for it. The problem is that the cost may be greater than the benefit. If workers are forced to take a fringe because it is "important," then they could be forced to pay more for something than it is worth to them. We can't quite understand the logic of forcing people to "buy" something that they do not believe is worth the cost. There are lots of things that people think are important for other people to have. But typically it is best to let individuals decide for themselves how much of these important things they buy with their own money. Individuals have information on their own preferences and circumstances that others do not know, and cannot know.

Critics might like to think that employers would pay for any given benefit. If the analysis of this section has led to any clear conclusion, it is that that the workers pay for what they get. They may not hand over a check for the benefits, but they give up the money nonetheless, through a reduction in their pay. If workers didn't give up anything for the fringe, we would have to conclude that the benefit was not worth anything to the workers, the supply curve would not move out, and the wage rate would not fall. That would mean that the employers would have to cover the full cost of the fringe, which would put them in the rather irrational position of adding to their costs without getting anything for it. Workers should not want that to happen if for no other reason than their job security would be threatened.

But critics might argue that managers don't know that certain fringes are "good" for business and their workers. That is often the case, and the history of business is strewn with the corpses of firms that failed to serve the interests of their workers and customers and who were forced into bankruptcy by other firms who were better at finding the best combination of fringes. We see the market as a powerful, though imperfect, educational system. If the critics know better than existing firms, they could make lots of money by pointing out to firms why they are

wrong and how they could make money from their employees by providing (selling) fringes not now being provided, or adjusting the combination of existing fringes in marginal ways.

We also don't believe that managers are the only ones who should search for the right combination of fringes. Workers should have an interest in joining the search, because they can gain in spite of the fact that their efforts will include a search for how their firms can make more money off them. If workers want more of one benefit, it would seem that all they would have to do is tell their bosses and show them how additional profits can be made *from the workers*. Workers, however, who want benefits without paying for them shouldn't waste their bosses' time. Managers hear from a lot of people who want something for nothing.

We think that workers and owners should talk as frankly about fringe benefits as they do about their wages. Workers earn their wages. The same is true for fringes. There's no gift involved. Both wages and fringes represent mutually beneficial exchanges between workers and their firms.

MANAGER'S CORNER: **Why Some Firms Pay for Their Employees' MBAs**

Education is sometimes said to be a good that stands in contrast with our road example. An educated person can provide others with whom he or she interacts with benefits. Lee and McKenzie can work together on this book in part because the other is "educated," meaning at its most fundamental level each can write and read what the other writes. Each benefit from the other's education, but neither contributed *directly* to the other's education expenses. One argument for government subsidies for education has been that because people who acquire education don't garner all the benefits from their education, then they will buy "too little" education, or extend their education only so long as their *personal* benefits were greater than their *personal* cost, which could mean that without the ability to communicate, much productive work would not be done.

This argument may hold for elementary and high school education, where the development of basic literacy is important, but it may not hold at the MBA level when practically all of the benefits seem to be *private*, meaning received by identified people, not *public*, meaning received by everyone in the broader community.

In this "Manager's Corner," we can extend our use of economic thinking to understand why firms behave the way they do. We start by noting that firms pay for some things for their workers but not other things. Why? We consider here an employee expense – an MBA education – that is sometimes covered and sometimes not covered by firms (consider the people in class). We also note that there is good reason to think that either the students or their firms should pay for the MBA education; the benefits are captured by the two groups. Our examination of these issues will help us draw out underlying principles, and the incentives that go with employer coverage of other work-related expenditures, not just education.

We suspect that many readers have a personal interest in this "Manager's Corner," given that they may be contemplating getting an MBA or some other advanced business degree, and hoping their employers will cover the cost. Why would any firm train its workers at the firm's expense?" The most general answer to that question is the same as the one given for why firms provide any fringe benefit: Firms make available some forms of training because, by doing so, they can make money off their workers. Training enables employers to increase worker productivity, to expand the supply of labor -- and to lower their wage bills. Employers sometimes offer training because their training cost is lower than the price the workers would have to pay if they got the training on their own. In such cases, employees gain by "buying" their training from their firm by way of reduced wages. However, an important theme of this "way of economic thinking" is that employer-financed training is no gift; it is a mutually beneficial trade between employers and employees. Of course, there are more details to be added to those generalities.

Firms cannot usually avoid providing some training for their workers, given that all workers must understand what is expected of them in their particular work environments. Workers must learn their companies' "culture," lines of communication, and the division of decision-making authority. However, such observations on training hide the full complexity of the decision relating to whom -- the firm or worker -- should be expected to pay for it. In almost all work environments, the costs of the training are usually divided, which raises an interesting question: *Along what conceptual lines should we expect the training costs to be divided?*[55] To the student-reader, the relevant question is "When can I expect my boss to cover the cost of my education?" The employer-reader sees the issue differently: "When should I cover the cost of my workers' education and, at the same time, avoid wasting money and sending the wrong incentive signals to my workers?"

Many workers -- including many skilled craftsmen (plumbers and carpenters, for example) -- pay for their own training.[56] Just about all undergraduate students and many, if not most, MBAs cover their own educational expenses. Many readers of this book know, through personal experience, that MBA students often pay for their graduate education while they are still employed by their firms. At the same time, some firms pay the tuition and fees of their managers who go back to college for MBAs. Again, what divides the two groups, those workers who train themselves and those who don't?

We suggest that the division is, to an important degree, based on the nature of the **human capital** that is acquired. Human capital is the accumulated skills and knowledge of

---

[55]By asking the question as we have, some readers might forget that education is a "good" that must in part be paid by the one receiving it. This is because a major part of the cost of education is the time devoted to study and class attendance. While students might be compensated for their time, they cannot avoid incurring the time cost.

[56]Skilled workers often pay for their training indirectly, by taking an apprenticeship with experienced craftsmen, which pays less than the workers could have received in some other job that does not provide training and the promise of a higher future income.

workers. If the acquired human capital is "specific" -- that is, the acquired skills are related to the particular needs of the worker's firm, which means that a worker with the acquired skills is not more attractive to other firms than any other worker (**specific human capital**) -- then the training will tend to be paid by the employer. The only reason the worker might cover the cost of such training for the development of specific human capital is that he or she might be promised a higher future income stream with the firm, and the present value of the additional income must be at least equal to the cost of the training.

However, the worker will rightfully fear that once he or she has incurred the training cost, the firm will renege on its part of the bargain. The fear can be especially relevant when the firm is financially unsound. Hence, the trained worker will be left without compensation for the cost incurred. The source of the basic problem is one that we have encountered before: *credibility*. Employers will tend to pay for the training involving specific human capital when their promise to repay workers in the future (through higher wages) for the costs the workers incur is not always credible, or believable.

Of course, when the employer's promise is tolerably credible, workers may actually cover the costs for their own firm-specific education (for example, they may study the personnel manual or product manuals on their own time). Workers will most likely cover such costs when they have been with their firms for a significant period of time and when the managers have a reputation for keeping their word.

Workers might also cover the costs of firm-specific training when workers can retaliate (at little expected cost) against their employers in the event that the employers renege on their agreements. For example, workers who use highly fragile pieces of test equipment might pay for their specific human capital, given that they can, with a low probability of detection, misuse or abuse the equipment under their control. In this case, the equipment can be viewed as the employer's bond. By putting workers in charge of equipment, the employer says, "If I ever fail to hold to my word, you can impose a substantial cost on me, perhaps more cost than I can impose on you." In such cases, employers should have no trouble getting their workers to do double time learning their jobs.

However, even in such cases, the problem of credibility does not evaporate. The workers' implied threat of destroying equipment must be believable. The more believable the threat, the more likely that the costs of firm-specific training can be incurred by the workers.[57] And in order for the threat to be believable, the worker must be able to impose costs on the employer without being caught, fired, and prosecuted. This leads to the interesting conclusion that if workers in charge of fragile equipment can be "caught" misusing and abusing that equipment, the employers will more likely have to cover the cost of their training. The worker's threat of retaliation will not be as forceful.

---

[57]The problem is really one of threat and counter-threat because the employer can also threaten to retaliate against the worker who retaliates for any failure to keep prior agreements.

Nevertheless, we might expect employers to pay for specific human capital when they have a reputation for fair and honest dealing. As we have argued before, employers are likely to be less risk averse than their employees, given that they may know more than their workers about how the workers' human capital will be utilized in the future. Employers can also spread the risk of the human capital investment over a large number of workers. By paying for their workers' specific human capital, employers can also reduce the employment risks of their workers. The training can be a way of saying to the workers: "We intend to keep you around for a while. Otherwise, we would not be investing in your skills. Once we give you the firm-specific training, you will be more valuable to us."

As a consequence, if the worker pays for the specific human capital, the employer would have to provide the worker with compensation that would have to include a risk premium, a cost that can be totally avoided by the employers who cover the training costs. Moreover, with the employers' heightened commitment to their workers' future employment, the workers should be expected to work for less than otherwise.

If the human capital is "general" -- that is, the acquired education and skills are wanted by a number of firms and therefore carry a market value for workers (**general human capital**) -- then the workers will tend to pay for the training themselves. The reasoning is much the same as the above, aside for the fact that the positions of the employers and employees are reversed. Employers will, understandably, be reluctant to pay for this type of training because the worker can then take the training and run. The workers will be in greater demand by the market, which means that they can, after receiving the training, be hired elsewhere at a higher wage, which reflects the market value of the acquired general human capital. Other firms will, consequently, hold back on training their workers, given that they can hire the trained workers from other firms without incurring the training costs. The firms that provide the training can see their market share erode as their more savvy competitors underprice them.

Hence, when all firms resist providing the training but pay higher prevailing market wages for those workers with the training (for example, graduate degrees in business), workers will voluntarily secure the training. Their higher expected lifetime earnings will cover their training costs.

Again, the basic problem in the covering of the costs is one of credibility, but this time it is the workers' credibility that is at stake. If workers can, in some way, assure their employers that they will remain with the firms after receiving the general human capital, then the firm will most likely cover the training costs. The costs can, in effect, be repaid by the workers by way of a lower-than-market wage for some time into the future.

Workers can enhance the credibility of their commitments in a number of ways. They can, through years of service to the firm, develop a reputation with their employers that their word is their bond. Workers can also, as a part of their pay packages, have some of their compensation deferred until, for example, retirement. The workers can also agree to lose some

or all of the deferred income if they decide to leave the company. The deferred income becomes, in effect, their bond, which is cashed in by their employer if the worker succumbs to the temptation of higher market wages and reneges on the training agreement. Here, naturally, the present discounted value of the deferred income that is subject to being lost by the workers must be greater than the cost of the general human capital they develop at the employer's expense.

Of course, workers can make formal contracts with their employers, which include a requirement that the worker stays with the firm for some specified number of years or else the worker will repay the entire cost of the training, and that the employee will not go into competition with his or her employer for some specified number of years.

One of the authors of this book got his Ph.D. funded in part by his first university employer (to the tune of half of his previous years' annual salary). However, he had to agree to stay with the university for two years for each year of graduate support. H&R Block, the tax preparation service, provides extensive training on the tax laws for its tax preparers, but it also requires them to agree not to go into the tax business outside H&R Block for several years.

Many MBA students who are reading this book as a part of a course assignment are probably having their graduate education paid for by their employers. That may seem odd, given that most MBA degrees increase the marketability and pay of graduates, which might be a problem for employers who are paying the bills. Our logic leads us to believe that those students will tend to have the following characteristics:

- First, the students whose employers are paying their educational tabs are probably older students who have been with their companies for a number of years. They have achieved some credibility with their employers, meaning their promise to stay with the firm carries weight.
- Second, it may also be that those students have won what is, in effect, a "prize" in an ongoing "tournament" organized by their employers. The educational prize has been designed to increase all worker productivity in the firm. In cases in which employer-paid general human capital is the result of a tournament, the employer would not necessarily be upset if the new MBAs leave the firm. The education could have still been a paying proposition because the firm has already been compensated for the cost of the MBAs by greater worker productivity.
- Third, a number of MBA students have probably signed some document with their firm that carries the weight of a contract and binds them to their firms for several years or requires them to repay the cost of their MBAs. Those students may have also agreed to repay the cost of those courses in which their performance does not meet some predetermined standard (for example, the students must receive a grade higher than a B). After all, the employer will want to make sure that the worker/students are no more

predisposed to shirk in the classroom than they are on the job. By having the grade restrictions, the employer will ensure that the education has the potential of paying off.

- Fourth, the students are in managerial positions in which the benefits of their having an MBA have the promise of showing up fairly rapidly in greater firm returns. The shorter the recovery period, the more likely the firm will cover the cost of managers' MBAs.
- Fifth, some of the students will have permitted a portion of their past compensation to be deferred to some point in the future, which can act like a bond. More generally, the employer will tend to select those workers for general education, like an MBA, who will incur a cost if they leave the firm.
- Sixth, the students will tend to come from the ranks of those who are on the executive "fast track," or have a great deal of promise in moving up the corporate ladder within the firm. Employers have a natural interest in making sure that such fast moving executives are well educated for their future posts. However, there is another complimentary reason for their selection for MBA programs. If the "fast-track" students leave their firms upon graduation, they will give up their expected higher status and income streams within the firm.
- Finally, students will also likely come from companies that have a promise of being around for a number of years. Financially shaky firms in highly unstable markets are going to be reluctant to pay for the cost of their workers' MBAs. Credit will, for them, be hard to come by. They will want their employees to use their own credit for their education, thereby freeing up the company's credit to finance company-specific investments in which the workers would not invest. Financially shaky companies will also not be able to count on being around to collect on the benefits of their workers' training. The workers in such companies will not likely have accepted much of their income in deferred forms and will not likely have strong expectations of a long career with their companies, factors that reinforce the tendency of workers to pay for their own MBAs.

All in all, we would expect, as a rule, most of the students whose education costs are covered by their employers to be weighted toward heavily experienced managers who work for established, stable, and generally large firms.

However, we hasten to add that it is only a manner of speaking when we say that employers will cover the cost of their workers' general human capital. In one way or another, we would expect workers to cover the cost, directly or indirectly. Firms that offer to fund the general education of their workers can expect to see, as a consequence, a greater supply of more qualified workers and a total wage bill that is lower than it would otherwise be.

Much training is, admittedly, a mix of firm-specific and general human capital components. All we can say is that the cost will tend to be divided according to whom -- the employer or employee -- benefits. The more firm specific the training, the greater the share of

the cost will be borne by the employers. Nothing is free in business, especially education. No matter what the form, someone will pay the piper.

Concluding Comments

Our message in this chapter and repeated elsewhere in this textbook, repeated and reinforced with analysis and anecdotes is simple: *Incentives are important*. They are worthy of serious reflection. But that doesn't mean to suggest that incentives are *all* that matter. Surely, many things matter. As noted earlier, leadership, product design, and customer service, as well as company adaptability, culture, and goals, also matter. However, we suspect that all of those good things in business might not matter very much or for long if the incentives are not right. In their effort to get incentives right, it is altogether understandable why some firms will cover the cost of the MBA degree program for some of their workers (and not others). In general, firms can be expected to cover the cost of an MBA when they, the firms, can expect to capture the benefits. On the other hand, the workers themselves can be expected to pay for their own degree expenses when they, the workers, expect to capture the benefits.

We hope our discussion of the importance of incentives in understanding the organization and performance of firms serves as an *incentive* to spend more time thinking and reading about incentives, a subject to which we will return later in the book and course.

Review Questions

1. Why are some firms "large" and other firms "small"? Use the concepts of "coordinating costs" in your answer
2. Suppose firms get smaller. Why might that happen?
3. If worker-monitoring costs go down, what will happen to the size of the firm?
4. What have been the various effects of the computer/telecommunication revolution on the sizes of firms?
5. Why would a firm hire its own accountants to keep the books but, at the same time, use outside lawyers to do its legal work?
6. If your firm fears being "held up" by an outside supplier of a critical part to your production process, what can your firm do to reduce the chance of a hold up?

CHAPTER 7

Market Failures: External Costs and Benefits

In its broadest definitional sense, collective action is the enactment and enforcement of law. The justification for all collective action, for government, lies in its ability to make men better off. This is where any discussion of the bases for collective action must begin.

James Buchanan

How much should government involve itself in the marketplace? How much does business want government involvement." These questions touch on one of the most important economic issues of our time: the division of responsibility between the public and private sectors. In general, economic principles would suggest that government undertake only functions that it can perform more efficiently than the market. As we will see, businesses are not always opposed to government involvement in the economy. Indeed, many businesses have incentives to try to make sure that government is more involved in the economy than is "efficient."

Economics provides a method for evaluating the relative efficiency of government and the marketplace. It enables the United States to identify which goods and services the market will fail to produce altogether, and which it will produce inefficiently. We saw in an earlier chapter that such market failures have three sources: monopoly power, external costs, and external benefits. Now, using the principles and graphic analyses developed in earlier chapters, we will take a closer look at external costs and benefits and at government attempts to capture them and correct market failures. (See later chapters on **monopoly** and **monopsony power**.)

External Costs and Benefits, Again

In a competitive market, producers must minimize their production costs in order to lower their prices, increase their production levels, and improve the quality of their products. Consumers must demonstrate how much they will pay for a product, and in what amount they will buy it. In a competitive market, production will move toward the intersection of the market supply and demand curves -- Q_1 in Figure 7.1. At that point the marginal cost of the last unit produced will equal its marginal benefit to consumers.

To the extent that the market moves toward equilibrium in supply and demand, it is efficient in a very special sense. As long as the marginal benefit of anything people do is greater than the marginal cost, people are presumed to be better off if quantity increases. In Figure 7.1, for each loaf of bread up to Q_1, the marginal benefit of consumption (as

shown by the demand curve) exceeds the marginal cost of production (as shown by the supply curve). Because the marginal cost of a loaf of bread is the value of the most attractive alternative forgone, people must be getting more value out of each of those loaves than they could from any alternative good. By producing exactly Q_1 loaves—no more and no less—the market extracts the possible surplus or excess benefits from production (see shaded area on the graph) and divides them among buyers and sellers. In this sense, production and distribution of economic resources can be said to be efficient.

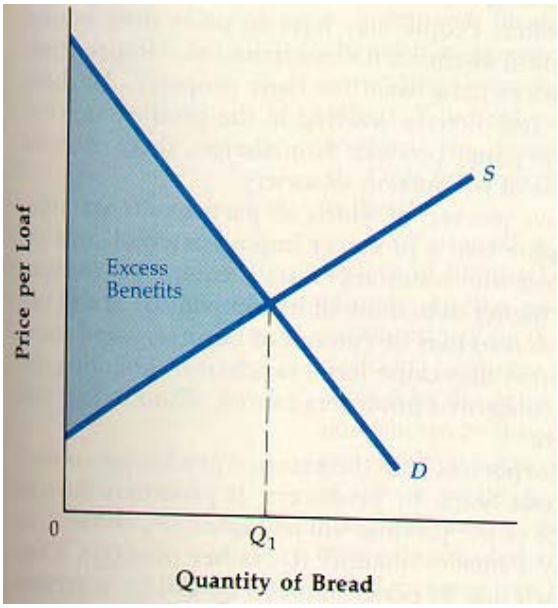


Figure 7.1 Marginal Benefit versus Marginal Cost

The demand curve reflects the marginal benefits of each loaf of bread produced. The supply curve reflects the marginal cost of producing each loaf. For each loaf of bread up to Q_1, the marginal benefits exceed the marginal cost. The shaded area shows the maximum welfare that can be gained from the production of bread. When the market is at equilibrium (when supply equals demand), all those benefits will be realized.

These results cannot be achieved unless competition is intense, buyers receive all the product's benefits, and producers pay all the costs of production. If such optimum conditions are not achieved, the market fails. Part of the excess benefits shown by the shaded area in the figure will not be realized by either buyers or sellers.

When exchanges between buyers and sellers affect people who are not directly involved in the trades, they are said to have external effects, or to generate externalities. **Externalities** are the positive or negative effects that exchanges may have on people who are not in the market. They are third-party effects. When such effects are pleasurable they are called external benefits. When they are unpleasant, or impose a cost on people other than the buyers or sellers, they are called external costs. The effects of external costs and benefits on production and market efficiency can be seen with the aid of supply and demand curves.

External Costs

Figure 7.2 represents the market for a paper product. The market demand curve, *D*, indicates the benefits consumers receive from the product. To make paper, the producers must pay the costs of labor, chemicals, and pulpwood. The industry supply curve, S_1, shows the cost on which paper manufacturers must base their production decisions. In a

perfectly competitive market, the quantity of the paper product that is bought will be Q_2, and the price paid by consumers will be P_1.

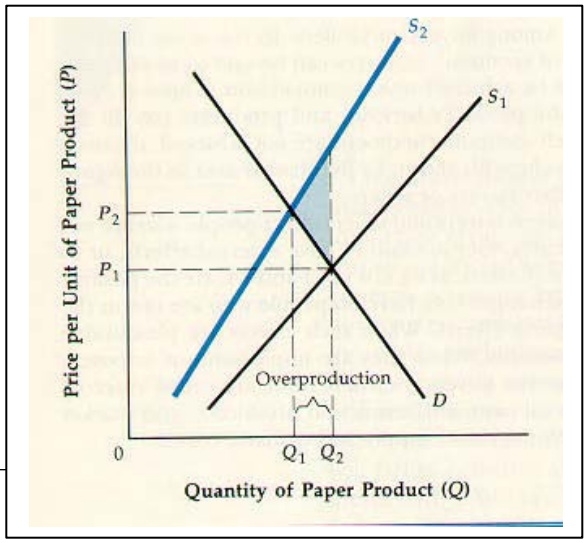


FIGURE 7.2 External Costs

Ignoring the external costs associated with the manufacture of paper products, firms will base their production and pricing decisions on supply curve S_1. If they consider external costs, such as the cost of pollution, they would operate on the basis of supply curve S_2, producing Q_1 instead of Q_2 units. The shaded area shows the amount by which the marginal cost of production of Q_2 -- Q_1 units exceeds the marginal benefits to consumers. It indicates the inefficiency of the private market when external costs are not borne by producers.

Producers may not bear all the costs associated with production, however. A by-product of the production process may be solid or gaseous waste dumped into rivers or emitted into the atmosphere. The stench of production may pervade the surrounding community. Towns located downstream may have to clean up the water. People may have to paint their houses more frequently or seek medical attention for eye irritation. Homeowners may have to accept lower prices than usual for their property. All these costs are imposed on people not directly involved in the production, consumption, or exchange of the paper product. Nonetheless, these external costs are part of the total cost of production to society.

In a perfectly competitive market, in which all participants act independently, survival may require that a producer impose external costs on others. An individual producer who voluntarily installs equipment to clean up pollution will incur costs higher than those of its competitors. It will not be able to match price cuts, and so in the long run may be out of business -- and some producers may not care whether they cause harm to others by polluting the environment. Even socially concerned producers cannot afford to care too much about the environment.

The supply curve S_2 incorporates both the external production costs of pollution and the private costs borne by producers. If producers have to bear all those costs, the price of the product will be higher (P_2 rather than P_1), and consumers will buy a small quantity (Q_1 rather than Q_2). Thus the true marginal cost of each unit of paper between Q_1 and Q_2 is greater than the marginal benefit to consumers. If consumers have to pay for external costs, they will value other goods more highly than those units. In a sense, then, the paper manufacturers are overproducing, by Q_2 -- Q_1 units. The marginal cost of those units exceeds their marginal benefit by the shaded triangular area.

Other examples of external costs that encourage overproduction are the highway congestion created by automobiles and the noise created by airplanes in and around airports. The argument can also be extended to include less obvious costs, like the death and destruction caused by speeding and reckless driving. If government does not penalize such negligent behaviors, people will produce them, at a potentially high external costs to others. In the same way, adult bookstores, X-rated movie houses, and massage parlors impose costs on neighboring businesses. Their sordid appearance drives away many people who might otherwise patronize legitimate businesses in the area.

External Benefits

Sometimes market inefficiencies are created by external benefits. Market demand does not always reflect all the benefits received from a good. Instead, people not directly involved in the production, consumption, or exchange of the good receive some of its benefits.

To see the effects of external benefits on the allocation of resources, consider the market for flu shots. The cost of producing vaccine includes labor, research and production equipment, materials, and transportation. Assuming that all those costs are borne by the producers, the market supply curve will be *S* in Figure 7.3.

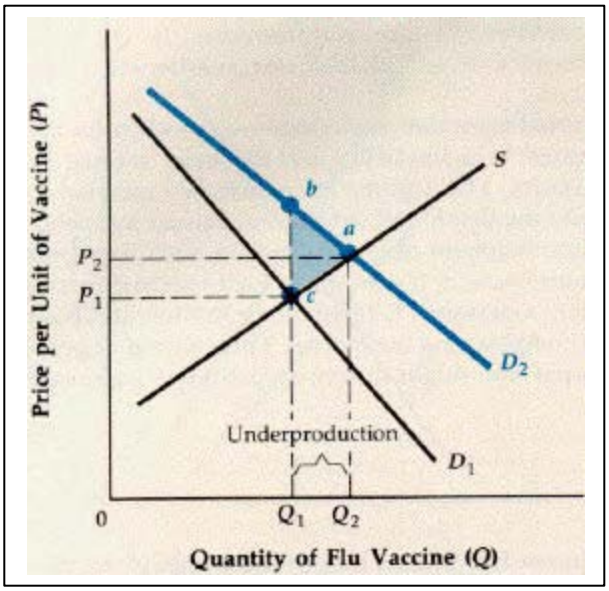


Figure 7.3 External Benefits

Ignoring the external benefits of getting flu shots, consumers will base their purchases on demand curve D_1 instead of D2. Fewer shots will be purchased than could be justified economically -- Q_1 instead of Q_2. Because the marginal benefit of each shot between Q_1 and Q_2 (as shown by demand curve D_2) exceeds it marginal cost of production, external benefits are not being realized. The shaded area *abc* indicates market inefficiency.

Individuals receive important personal benefits from flu shots. The fact that many millions of people pay for them every year shows that there is a demand, illustrated by curve D_1 in Figure 7.3. In getting shots for themselves, however, people also provide external benefits for others. By protecting themselves, they reduce the probability that the flu will spread to others. When others escape the medical expenses and lost work

time associated with flu, those benefits are not captured in the market demand curve, D_1. Only in the higher societal demand curve, labeled D_2, are those benefits realized.

Left to itself, a perfectly competitive market will produce at the intersection of the market supply and market demand curves (S and D_1), or at point *c*. At that point the equilibrium price will be P_1 and the quantity produced will be Q_1. If external benefits are considered in the production decision, however, the marginal benefit of flu shots between Q_1 and Q_2 (shown by the demand curve D_2) will exceed their marginal cost of production (shown by the supply curve). In other words, if all benefits, both private and external, were considered, Q_2 shots would be produced and purchased at a price of P_2. At Q_2, the marginal cost of the last shot would equal its marginal benefit. Social welfare would rise by an amount equal to the triangular shaded area *abc*.

Because a free market can fail to capture such external benefits, government action to subsidize flu shots may be justified. On such grounds governments all over the world have mounted programs to inoculate people against diseases like smallpox. The external benefits argument has also been used to justify government support of medical research. It can also be extended to services such as public transportation. City buses provide direct benefits to the general population. An informed and articulate citizenry raises both the level of public discourse and the general standard of living.[1] Public parks and environmental programs can also provide external benefits that are not likely to be realized privately, because of their high cost to individuals. Again, government action may be required to supplement private efforts.

The Pros and Cons of Government Action

More often than not, exchanges between buyers and sellers affect others. People buy clothes partly to keep warm in the winter and dry in the rain, but most people value the appearance of clothing at least as much as its comfort. We choose clothing because we want others to be pleased or impressed (or perhaps irritated). The same can be said about the cars we purchase, the places we go to eat, the records we buy, even the colleges we attend. We impose the external effects of our actions deliberately as well as accidentally.

The presence of externalities in economic transactions does not necessarily mean that government should intervene. First, the economic distortions created by externalities are often quite small, if not inconsequential. So far our examples of external costs and benefits involved possibly significant distortions of market forces. In Figure 7.4, however, the supply curve S_2, which incorporates both private and external costs, lies only slightly to the left of the market supply curve, S_1. The difference between the market output level, Q_2, and the optimum output level, Q_1, is small, as is the market inefficiency, shown by the shaded triangular area. Therefore little can be gained by government intervention.

---

[1] The ratio of public to private benefits varies by educational levels. Elementary school education develops crucial social and communication skills; its private benefits are virtually side effects. At the college level, however, the private benefits to students may dominate the public benefits. Thus elementary education is supported almost entirely by public sources, while college education is only partially subsidized.

This limited benefit must be weighed against the cost of government action. Whenever government intervenes in any situation, agencies are set up, employees are hired, papers are shuffled, and reports are filed. Almost invariably, suits are brought against firms and individuals who have violated government rules. In short, significant costs can be incurred in correcting small market inefficiencies. If the cost of government intervention exceeds the cost of the market's inefficiencies, government action will actually increase inefficiency.

A second reason for limiting government action is that it generates external costs of its own. If government dictates the construction methods to be used in building homes, the way mothers deliver their babies, or the hair lengths of government workers, the people who set the standards impose a cost—which may be external to them—on those who do not share their standards. We may agree with some government rules, but strenuously object to others. On balance, such government intervention is as likely to hurt us as help us.

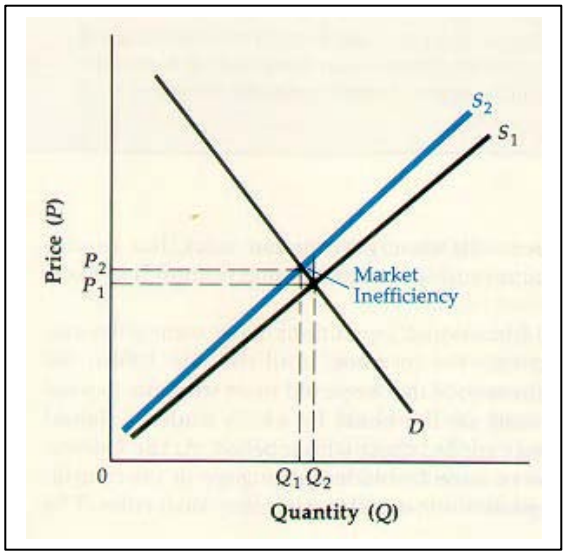


Figure 7.4 Is Government Action Justified?

Because of external costs, the market illustrated produces more than the efficient output. Market inefficiency, represented by the shaded triangular area, is quite small—so small that government intervention may not be justified on economic grounds alone.

Government dictates in educational institutions have sometimes imposed onerous costs on students. For instance, until the late 1960s, the University of Virginia had a dress code that required male students to wear coats and ties. Colleges routinely set the hours by which students should return to their dormitories and expelled those who rebelled. At the University of California, students were once forbidden to engage in on-campus political activity. Costs are imposed on those who must obey such rules. The more centralized the government that is setting the standards, the less opportunity people will have to escape the rules by moving elsewhere.

In certain markets, government action may not be necessary. Over the long run, some of the external costs and benefits that cause market distortions may be internalized. That is, they may become private costs and benefits. Suppose the development of a park would generate external benefits for all businesses in a shopping district. More customers would be attracted to the district, and more sales would be made. An alert entrepreneur could internalize those benefits by building a shopping mall with a park in

the middle. Because the mall would attract more customers than other shopping areas, the owner could benefit from higher rents. When shopping centers can internalize such externalities, economic efficiency will be enhanced—without government intervention.

When Walt Disney built Disneyland, he conferred benefits on merchants in the Anaheim area. Other businesses quickly moved in to take advantage of the external benefits –the crowds of visitors—spilling over from the amusement park. Disney did not make the same mistake twice. When he built Disney World in Orlando, he bought enough land so that most of the benefits of the amusement park would stay within the Disney domain. Inside the more than six thousand acres of Disney-owned land in Florida, development has been controlled and profits captured by the Disney Corporation. Although other businesses have established themselves on the perimeters of Disney World, their distance from its center makes it more difficult for them to capture external benefits from the amusement park.

Methods of Reducing Externalities

Government action can undoubtedly guarantee that certain goods and services will be produced more efficiently. The benefits of such action may be substantial, even when compared with the costs. In such cases, only the form of government intervention remains to be determined. Government action can take several forms; persuasion; assignment of communal property rights to individuals; government production of goods and services; regulation of production through published standards; and control of product prices through taxes, fines, and subsidies. Economists generally argue that if government is going to intervene, it should choose the least costly means sufficient for the task at hand.

Persuasion

External costs arise partly because we do not consider the welfare of others in our decisions. Indeed, if we fully recognized the adverse effects of our actions on others, external cost would not exist. Our production decisions would be based as much as possible on the total costs of production to society.

Thus government can alleviate market distortions by persuading citizens to consider how their behavior affects others. Forest Service advertisements urge people not to litter or to risk forest fires when camping. Other government campaigns encourage people not to drive if they drink, to cultivate their land so as to minimize erosion, and to conserve water and gas. Although such efforts are limited in their effect, they may be more acceptable than other approaches, given political constraints.

Persuasion can take the form of publicity. The government can publish studies demonstrating that particular products or activities have external costs or benefits. The resultant publicity may in turn encourage those activities with external benefits and discourage those activities with external costs. The government has, for example, used this method in the case of cigarettes, publishing studies showing the external costs of smoking.

Assignment of Property Rights

As we saw in Chapter 1, when property rights are held communally or left unassigned, property tends to be overused. As long as no one else is already using the property, anyone can use it without paying for its use. Costs that are not borne by users, of course, are passed on to others as external costs. When public land was open to grazing in the West 150 years ago, for instance, ranchers allowed their herds to overgraze. The external cost of their indiscriminate use of the land has been borne by later generations, who have inherited a barren, wasted environment.

Thus the assignment of property rights can eliminate some externalities. If land rights are assigned to individuals, they will bear the cost of their own neglect. If owners allow their cattle to strip a range of its grass, they will no longer be able to raise their cattle there—and the price of the land will decline with its productivity.

Some resources, such as air and water, cannot always be divided into parcels. In those cases, the property rights solution will work poorly, if at all.

Government Production

Through nationalization of some industries, government can attempt to internalize external costs. The argument is that because government is concerned with social consequences, it will consider the total costs of production, both internal and external. On the basis of that argument, governments in the United States operate schools, public health services, national and state parks, transportation systems, harbors, and electric power plants. In other nations, government also operates major industries, such as the steel and automobile industries.

Government production can be a mixed blessing. When other producers remain in the market, government participation may increase competition. Sometimes it means the elimination of competition. Consider the U.S. Postal Service, which has exclusive rights to the delivery of first-class mail. As a government agency, the Post Office is not permitted to make a profit that can be turnover to shareholders. Because of its market position with little competition for home delivery of mail, however, it may tolerate higher costs and lower work standards than competitive firms.

Some government production, such as the provision of public goods like national defense, is unavoidable. In most cases, however, direct ownership and production may not be necessary. Instead of producing goods with which externalities are associated, government could simply contract with private firms for the business. That is precisely how most states handle road construction, how several states handle the penal system, and how a few city governments provide ambulance, police, and firefighting services.

Taxes and Subsidies

Government can deal with some external costs by taxing producers. Pollution can be discouraged by a tax on either the pollution itself or the final product. Taxing the

pollution emitted by firms internalizes external costs, increasing total costs to the producer. Imposing such taxes should have a twofold effect in reducing pollution. First, many producers would find the cost of pollution control cheaper than the pollution tax. Second, the tax would raise the prices of final products, reducing the number of units consumed -- and hence reducing the level of pollution.

The size of the tax can be adjusted to achieve whatever level of pollution is judged acceptable. If a tax on $1 per unit produced does not reduce pollution sufficiently, the tax can be raised to $2. In terms of Figure 7.2, the ideal tax would be just enough to encourage producers to view their supply curve as S_2 instead of S_1. The resulting cutback in production from Q_2 to Q_1 would eliminate market inefficiency, represented by the shaded area *abc*.

Theoretically, the government could achieve the same result by subsidizing firms in their efforts to eliminate pollution. It could give tax credits for the installation of pollution controls or pay firms outright to install the equipment. In fact, until 1985, the federal government used tax credits to encourage the installation of fuel-saving devices, which indirectly reduced pollution.

Production Standards

Alternatively, the government could simply impose standards on all producers. It could rule, for example, that polluters may not emit more than a certain amount of pollutants during a given period. Offenders would either have to pay for a cleanup or risk a fine. A firm that flagrantly violated the standard might be forced to shut down.

Choosing the Most Efficient Remedy for Externalities

Selecting the most efficient method of minimizing externalities can be a complicated process. To illustrate, we will compare the costs of two approaches to controlling pollution, government standards versus property rights

Suppose five firms are emitting sulfur dioxide, a pollutant that causes acid rain. The reduction of the unwanted emissions can be thought of as an economic good whose production involves a cost. We can assume that the marginal cost of reducing sulfur dioxide emissions will rise as more and more units are eliminated. We can also assume that such costs will differ from firm to firm. Table 7.1 incorporates these assumptions. Firm A, for example, must pay $100 to eliminate the first unit of sulfur dioxide and $200 to eliminate the second. Firm B must pay $200 for the first unit and $600 for the second. Although the information in the table is hypothetical, it reflects the structure of real-world pollution clean-up costs. The technological fact of increasing marginal costs faces firms when they clean up the air as well as when they produce goods and services.

Suppose the Environmental Protection Agency (EPA) decides that the maximum acceptable level of sulfur dioxide is ten units. To achieve that level, the EPA prohibits firms from emitting more than two units of sulfur dioxide each. If each firm were emitting five units, each would have to reduce its emissions by three units. The total cost

of meeting the limit of two units is shown in the lower half of Table 7.1. Firm A incurs the relatively modest cost of $700 ($100 + $200 + $400). But firm B must pay $2,600 ($200 + $600 + $1,800). The total cost to all firms is $13,500.

What if the EPA adopts a different strategy and sells rights to pollute? Such rights can be thought of as tickets that authorize firms to dump a unit of waste into the atmosphere. The more tickets a firm purchases, the more waste it can dump, and the more cleanup costs it can avoid.

TABLE 7.1 Costs of Reducing Sulfur Dioxide Emissions

	A	B	C	D	E
Marginal cost of eliminating each unit of pollution:					
First unit	$ 100	$ 200	$ 200	$ 600	$1,000
Second unit	200	600	400	1,000	2,000
Third unit	400	1,800	600	1,400	3,000
Fourth unit	800	5,400	800	1,800	4,000
Fifth unit	1,600	16,200	1,000	2,200	5,000

Cost of Reducing Pollution by Establishment of Government Standards		*Cost of Reducing Pollution by Sale of Pollution Rights*	
Cost to A of eliminating 3 units	$ 700	Cost to A of eliminating 4 units	$1,500
Cost to B of eliminating 3 units	2,600	Cost to B of eliminating 2 units	800
Cost to C of eliminating 3 units	1,200	Cost to C of eliminating 5 units	3,000
Cost to D of eliminating 3 units	3,000	Cost to D of eliminating 3 units	3,000
Cost to E of eliminating 3 units	6,000	Cost to E of eliminating 1 unit	1,000
Total cost of five units	**$13,500**	**Total cost of five units**	**$9,300**

Remember that the EPA can control the number of tickets it sells. To limit pollution to the maximum acceptable level of ten units, all it needs to do is sell no more than ten tickets. Either way, whether by pollution standards or rights, the level of pollution is kept down to ten units, but the pollution rights method allows firms that want to avoid the cost of a cleanup to bid for tickets.

The potential market for such rights can be illustrated by conventional supply and demand curves, as in Figure 7.5. The supply curve is determined by EPA policymakers, who limit the number of tickets to ten. Because in this example the supply is fixed, the supply curve must be vertical (perfectly inelastic). Whatever the price, the number of pollution rights remains the same. The demand curve is derived from the costs firms must bear to clean up their emissions. The higher the cost of the cleanup, the more

attractive pollution rights will be. As with all demand curves, price and quantity are inversely related. The lower the price of pollution rights, the higher the quantity demanded.

Figure 7.5 Market for Pollution Rights

Reducing pollution is costly (see Table 7.1). It adds to the costs of production, increasing product prices and reducing the quantities of products demanded. Therefore firms have a demand for the right to void pollution abatement costs. The lower the price of such rights, the greater the quantity of rights that firms will demand (see Table 18,2). If the government fixes the supply of rights at ten and sells those ten rights to the highest bidders, the price of the rights will settle at the intersection of the supply and demand curves -- here, $1,500.

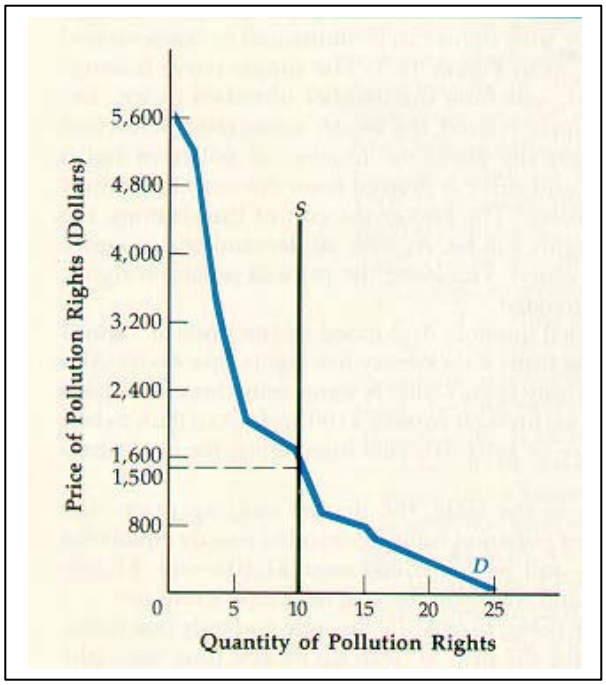


Table 7.2 shows the total quantity demanded by the firms at various prices. At a price of zero, the firms want twenty-five rights (five each). At a price of $201, they demand only twenty-one. A wants only three, for it will cost less to clean up its first two units (at costs of $100 and $200) than to buy rights to emit them at a price of $201. B wants four rights, for its cleanup costs are higher.

Given the information in the table, the market clearing price—the price at which the quantity of property rights demanded exactly equals the number of rights for sale—will be something over $1,400—say $1,500. Who will buy those rights, and what will the cost of the program be?

At a price of $1,500 per ticket, firm A will buy one and only one ticket. At that price, it is cheaper for the firm to clean up its first four units (the cost of the cleanup is $100 + $200 + $400 + $800). Only the fifth unit, which would cost $1,600 to clean up, makes the purchase of a $1,500 ticket worthwhile. Similarly, firm B will buy three tickets, firm C none, firm D two, and firm E four.

The cost of any cleanup must be measured by the value of the resources that go into it. The value of the resources is approximated by the firm's expenditures on the cleanup—not by their expenditures on pollution tickets. (The tickets do not represent real resources, but a transfer of purchasing power from the firms to the government.) Accordingly, the economic cost of reducing pollution to ten units is $9,300; $1,500 for firm A. $800 for B, $3,000 each for C and D, and $1,000 for E. This figure is

significantly less than the $13,500 cost of the cleanup when each firm is required to eliminate three units of pollution. Yet in each case, fifteen units are eliminated. In shirt, the pricing system is more economical—more cost-effective or efficient—than setting standards. Because it is more efficient, it is also the more economical way of producing goods and services. More resources go into production and less into cleanup.

TABLE 7.2 Demand for Property Rights

Price	Quantity	Price	Quantity
$ 0	25	$1,601	9
101	24	1,801	7
201	21	2,001	6
401	19	2,201	5
601	16	3,001	4
801	14	4,001	3
1,001	11	5,001	2
1,401	10	5,601	0

The idea of selling rights to pollute may not sound attractive, but it makes sense economically. When the government sets standards, it is giving away rights to pollute. In our example, telling each firm that it must reduce its sulfur dioxide emissions by three units is effectively giving them each permission to dump two units into the atmosphere. One might ask whether the government should be giving away rights to the atmosphere, which has many other uses besides the absorption of pollution. Though some pollution may be necessary to continued production, that is no argument for giving away pollution rights. Land is needed in may production processes, but the Forest Service does not give away the rights to public lands. When pollution rights are sold, on the other hand, potential users can express the relative values they place on the right to pollute.[2] In that way, rights can be assigned to their most valuable and productive uses.

MANAGER'S CORNER: **How Honesty Pays in Business**

There exist the popular perception that markets fail because business is full of dishonest scoundrels – especially high ranking executives -- who cheat, lie, steal, and worse to increase their profits. This perception is reflected in and reinforced by the way business people are depicted in the media. According to one study, during the 1980s almost 90 percent of all business characters on television were portrayed as corrupt.[3] No one can

[2] Note that the system allows environmental group as well as producers to express the value they place on property rights. If environmental groups think ten units of sulfur dioxide is too much pollution, they can buy some of the tickets themselves and then not exercise their right to pollute.

[3] See page 146 of Robert Lichter, Linda Lichter, and Stanley Rothman, Watching America (New York: Prentice Hall, 1990).

deny that people in business have done all kinds of nasty things for a buck. But the impression of pervasive dishonest business people is greatly exaggerated. Business people are no more likely to behave dishonestly than other people. In fact, there are reasons to believe that business people might be more honest than the typical American on the street. Moreover, there are ways business people can commit themselves to incentive arrangements that motivate honest behavior in ways that their customers find convincing.

The Role of Honesty in Business

The case to be made for honesty in business is not based on any claim that business people are particularly virtuous, or ethical to the core of their beings. We can make no claim to keen insights into the virtue of business people or anyone else. We might even be persuaded that business people have less virtue on average than do those who choose more caring occupations, such as teachers, social workers, missionaries, and nurses. But we do claim to know one simple fact about human behavior, and that is people respond to incentives in fairly predictable ways. In particular, the lower the personal cost of dishonesty, the greater the extent of dishonestly within most identified groups of people. If business people act honestly to an unusual degree (or different from what other people in other situations do), it must be in part because they expect to pay a high price for behaving dishonestly. This is, in fact, the case because business people have found, somewhat paradoxically, that they can increase profits by accepting institutional and contractual arrangements that impose large losses on them if they are dishonest.

Though seldom mentioned, most business activity requires a high degree of honest behavior. If business is going to be conducted at any but the simplest level, products must be represented honestly, promises must be kept, costly commitments must be made, and business people must cooperate with each other to take the interests of others, particularly consumers, into consideration. Indeed, if the proverbial man from Mars came down and observed business activity, he might very well conclude that business people are extraordinarily honest, trusting, and cooperative. They sell precious gems that really are precious to customers who cannot tell the difference between a diamond and cut glass. They promise not to raise the price of a product once customers make investments that make switching to another product costly, and they typically keep the promise. They make good faith pledges that the businesses they own, but are about to sell, will continue to give their customers good service. They commit themselves to costly investments to serve customers knowing the investments will become worthless if customers shift their business elsewhere.

The way business people behave in the marketplace suggests a level of morality that is at variance with the self-interest that economists assume, in their theoretical models, motivates business activity. Some argue that the economist's assumption of self interest is extreme, and we recognize that many people, including many business people, behave honestly simply because they feel it is the right thing to do. But few would recommend that we blindly trust in the honesty of others when engaged in business activity. The person who is foolish enough to assume that all business people are honest

and trustworthy only has to encounter a few who are not to find himself separated quickly from his wealth.

Is there a contradiction here between the honesty that characterizes most business activity and the fact that business people are not generally assumed to be honest? The answer is no. Indeed, the reason business people generally behave honestly is best explained by the fact that it would be foolish to assume that they are honest. And many business people are honest precisely because others assume they won't be.

It is easy to imagine a situation in which business people can profit at the expense of their customers, workers, and others with whom they deal if they behave deceitfully. For example, the quality of many products (say used cars or diamonds) is difficult for consumers to easily determine. The seller who takes advantage of this by charging a high quality price for a low quality product would capture extra profits from the sale. A business owner who is about to retire can profit by making promises not to be fulfilled until after his retirement, and which he does not plan to keep. The monopoly producer of a superior product (but one which requires the consumer to make costly investments in order to use it) can offer the product at a low price and then, once the consumer becomes dependent on it, increase the price significantly. Other examples of the potential profit from dishonest behavior are easily imagined. In fact, such examples are about the only type of behavior some people ever associate with business.

Again, we want to emphasize that dishonest behavior of the above type does occur. But such dishonest behavior is the exception, not the rule of much business, despite the story-telling talents of Hollywood writers. The reason is that in addition to being a virtue from a strictly moral perspective, honesty is also important for quite materialistic reasons. An economy in which people deal with each other honestly can produce more wealth than one in which people are chronically dishonest. So there are gains to be realized from honesty, and when there are gains to be captured there are people who, given the opportunities available in market economies, will devise ways to capture them.

A businessperson who attempts to profit from dishonest dealing faces the fact that few people are naively trusting. It may be possible to profit from dishonesty in the short run, but those who do so find it increasingly difficult to get people to deal with them in the long run. And in some businesses it is extremely difficult to profit from dishonesty even in the short run. How many people, for example, would pay full price for a "genuine" Rolex watch, or diamond necklace, from someone selling them out of a Volkswagen van at the curb of a busy street? Without being able to provide some assurance of honesty, the opportunities to profit in business are very limited.

So business people have a strong motivation to put themselves in situations in which dishonest behavior is penalized. Only by doing so can they provide potential customers, workers, and investors with the assurance of honest dealing required if they are to become actual customers, workers, and investors.

The advantage of honesty in business can be illustrated by considering the problem facing Mary who has a well-maintained 1990 Honda Accord that she is willing to sell for as little as $4,000. If interested buyers know how well maintained the car is,

they would be willing to pay as much as $5,000 for it. Therefore, it looks like it should be possible for a wealth-increasing exchange to take place since any price between $4,000 and $5,000 will result in the car being transferred to someone who values it more than the existing owner. But there is a problem. Many owners of 1990 Honda Accords who are selling their cars are doing so because their cars have not been well and are about to experience serious mechanical problems. More precisely, assume that that 75 percent of the 1990 Honda Accords being sold are in such poor condition that the most a fully informed buyer would be willing to pay for them is $3,000, with the other 25 percent worth $5,000. This means that a buyer with no information on the condition of a car for sale would expect a 1990 Honda Accord to be worth, on average, only $3,500. But if buyers are willing to pay $3,500 for a 1990 Accord, many of the sellers whose cars are in good condition will refuse to sell, as is the case with Mary who is unwilling to sell for less than $4,000.

So the mix of 1990 Accords for sale will tilt more in the direction of poorly maintained cars, their expected value will decline, and even fewer well-maintained 1990 Accords will be sold. This situation is often described as a market for "lemons," and illustrates the value of sellers being able to commit themselves to honesty. [4] If Mary could somehow convince potential buyers of her honesty when she claims her Accord is in good condition, she would be better off, and so would those who are looking for a good used car. The advantage of being able to commit to honesty in business extends to any situation where it is difficult for buyers to determine the quality of products they are buying.

The advantages of honesty in business and the problem of trying to provide credible assurances of that honesty can also be illustrated as a game. In Figure 7.6, we present a payoff matrix for a buyer and a seller giving the consequences from different choice combinations. The first number in the brackets gives the payoff to the seller and the second number gives the payoff to the buyer. If the seller is honest (the quality of the product is as high as he claims) and the buyer trusts the seller (she pays the high-quality price), then both realize a payoff of 100. On the other hand, if the seller is honest but the buyer does not trust him, then no exchange takes place and both receive a payoff of zero. If the seller is dishonest while the buyer is trusting, then the seller captures a payoff of 150, while the buyer gets the sucker's payoff of -50. Finally, if the seller is dishonest and the buyer does not trust him, then an exchange takes place with the buyer paying a low quality price but getting a lower quality product than she would be willing to pay for, with both the seller and buyer receiving a payoff of 25. From a joint perspective, honesty and trust are the best choices since this combination results in more wealth for the two to share. But this will not be the outcome, given the incentives created by the payoffs in Figure 7.6. The buyer will not trust the seller. The buyer knows that if her trust of the seller is taken for granted by the seller then he would attempt to capture the largest possible payoff from acting dishonestly. On the other hand, if he believes she does not trust him his highest payoff is still realized by acting dishonestly. So she will reasonably

[4]The general problem of "lemons" is discussed by George A. Akerlof, "The Market for Lemons: Qualitative Uncertainty and the Market Mechanism," Quarterly Journal of Economics, Vol. 84 (1970): 488-500.

expect the seller to act dishonestly. This is a self-fulfilling expectation since when the seller doesn't expect to be trusted, his best response is to act dishonestly.

Figure 7.6 The Problem of Trust in Business

		BUYER	
		Trust	**Doesn't Trust**
SELLER	**Honest**	(100, 100)	(0, 0)
	Dishonest	(150, -50)	(25, 25)

The seller would clearly be better off in this situation (and so would the buyer) if he somehow created an arrangement that reduced the payoff he could realize from acting dishonestly. If, for example, the seller arranged it so he received a payoff of only 50 from acting dishonestly when the buyer trusted him, as is shown in Figure 7.7, then the buyer (assuming she knows of the arrangement) can trust the seller to respond honestly to her commitment to buy. The seller's commitment to honesty allows both seller and buyer to each realize a payoff of 100 rather than the 25 they each receive without the commitment.

But how can a seller commit him or herself to honesty in a way that is convincing to buyers? What kind of arrangements can sellers establish that penalize them if they attempt to profit through dishonesty at the expense of customers?

There are many business arrangements, and practices, that can cause sellers to commit to honest dealings. We will briefly consider some of them here. The arrangements are varied, as one would expect, since the ways a seller could otherwise profit from dishonest activity are also varied.

Notice that our discussion of the situation described in Figure 7.6 implicitly assumes that the buyer and seller deal with each other only one time. This is clearly a situation in which the temptation for the seller to cheat the buyer is the strongest, since the immediate gain from dishonesty will not be offset by a loss of future business from a mistreated buyer. If a significant amount of repeat business is possible, then the temptation to cheat decreases, and may disappear. What the seller gains from dishonest dealing on the first sale can be more than offset by the loss of repeat sales. So, one way sellers can attempt to move from the situation described in Figure 7.6 to the one described in Figure 7.7 is by demonstrating that they are in business for the long run. For example, selling out of a permanent building with the seller's name or logo on it, rather than a Volkswagen van, informs potential customers that the seller has been (or plans on being) around for a long time. Sellers commonly advertise how long they have been in business (for example, "Since 1942" is added under the business name), to inform people that they have a history of honest dealing (or otherwise they would have been out of business long ago) and plan on remaining in business.

As we have seen, however, in our discussion of "the last period problem," the advantages motivated by repeated encounters tend to break down if it is known that the encounters will come to an end at a specified date. For this reason firms will attempt to maintain continuity beyond what would seem to be a natural end-period. Single proprietorships, for example, would seem to be less trustworthy when the owner is about to retire, or sell. But, as discussed earlier, a common way of reducing this problem is for the owner's offspring to join the business ("Samson and Sons" or "Delilah and Daughters") and ensure continuity after their parent's retirement. Indeed, even though large corporations have lives that extend far beyond that of any of their managers, they often depend on single proprietorships to represent and sell their products. As indicated earlier in the book with our example of Caterpillar, the heavy equipment company, it is common for such corporations to have programs to encourage the sons and daughters of these single proprietors to follow in their parents' footsteps.

Figure 7.7 The Problem of Trust in Business, Again

		BUYER	
		Trust	**Doesn't Trust**
SELLER	**Honest**	(100, 100)	(0, 0)
	Dishonest	(50, -50)	(25, 25)

The advantage of letting people know that you have been, and are planning to be, in business a long time is that it informs them that you have something to lose –potential future business -- if you engage in dishonest dealing. In effect, you are providing potential customers with a **hostage**, something of value that one party to a contract (the customer) can destroy if the other party (seller) does not keep its promises. There are numerous other ways that businesses create arrangements to provide hostages in ways that make their commitments to honest dealing credible. Before examining some of these arrangements, however, it is important to consider an important feature that hostages should have.

The use of hostages has a long history, and is traditionally thought of as a way to reduce the likelihood of hostilities between two countries or kingdoms. For example, if King A intended to wage war on Kingdom C and wanted to keep Kingdom B neutral, he could assure King B of his good faith by yielding up his beloved daughter to King B as a hostage. Assuming King A really did love his daughter, he would then be very reluctant to break his promise and invade Kingdom B after conquering Kingdom C. But even if King A does have a compelling incentive not to wage war against King B as long as his daughter is King B's hostage, a potential problem remains. King B may find the daughter so attractive that he values her more than her father's promise not to invade. Therefore, King B may decide to join with Kingdom C against King A and keep the daughter for himself. This suggests that an ugly daughter (one only a father could love!) makes a better hostage than a beautiful daughter.

The general proposition that comes from this example is that the best hostage is one that the person giving it up values highly and which the person receiving it values not at all. The example also suggests that sometimes it is best, particularly if the hostage is valuable to the person holding it, for the parties to exchange hostages. For example, if King A only has beautiful daughters then the best arrangement may be for him to exchange a beautiful daughter for one of King B's handsome sons (presumably for Queen A's keeping). Of course, it is now important that King B values his son more than he does King A's daughter and that Queen A values her daughter more than she does King B's son.

A firm's reputation can be thought of as a hostage that the firm puts in the hands of its customers as assurance that it is committed to honest dealing. A firm's reputation is an ideal hostage because it is valuable to the firm, but has no value to customers apart from its ability to ensure honesty. A firm has a motivation to remain honest in order to prevent its reputation from being destroyed by customer dissatisfaction, but customers cannot capture the value of the reputation for themselves. The more a firm can show that it values its reputation, the better hostage it makes.

Consider the value of a logo to a firm. Companies commonly spend what seems an enormous amount of money for logos to identify them to the public. Well-known artists are paid handsomely to produce designs that do not seem any more attractive than those that could be rendered by lesser-known artists (many of whose artistic efforts have never gone beyond bathroom walls). Furthermore, companies are seldom shy about publicizing the high costs of their logos.

It may seem wasteful for a company to spend so much for a logo, and silly to let consumers know about the waste (the cost of which ends up in the price of its products). But expensive logos make sense when we recognize that much of the value of a company's logo depends on its cost. The more expensive a company's logo, the more that company has to lose if it engages in business practices that harm its reputation with consumers, a reputation embodied in the company logo. The company that spends a lot on its logo is effectively giving consumers a hostage that is very valuable to the company. Consumers have no interest in the logo except as an indication of the company's commitment to honest dealing, but will not hesitate to destroy the value of the logo (hostage) if the company fails to live up to that commitment.

Expensive logos are an example of how businesses make non-salvageable investments to penalize themselves if they engage in dishonest dealing. Such investments are particularly common when the quality of the product is difficult for consumers to determine. The products sold in jewelry stores, for example, can vary tremendously and few consumers can judge that value themselves. Those jewelry stores that carry the more expensive products want to be convincing when they tell customers that those products are worth the prices being charged. One way of doing this is by selling jewelry in stores with expensive fixtures that would be difficult to use in other locations: ornate chandeliers, unusually shaped display cases, expensive counter tops, and generous floor space. What could the store do with this stuff if it went out of business? Not much, and this tells the customers that the store has a lot to lose by misrepresenting

its merchandise to capture short-run profits. Non-salvageable investments serve as hostages that sellers put into the hands of customers.

Another rather subtle way that sellers use "hostages" to provide assurances of honesty is by letting consumers know that they (the sellers) are making lots of money. If it is known that a business is making a lot more profit from its existing activity than it could make in alternative activities, consumers will have more confidence that the business won't risk that profit with misleading claims. The extra profits of the business are a hostage that will be destroyed by consumers' choices if the business begins employing dishonest practices.[5] Expensive logos and non-salvageable capital are not only hostages in themselves, they also inform consumers that the firm is making enough money to afford such extravagances. Expensive advertising campaigns, often using well-known celebrities, also serve the same purpose. Through expensive advertising, a company is doing more than informing potential customers about the availability of the product; it is letting them know that it has a lot of profits to lose by misrepresenting the quality of the product.[6]

The idea of firms intentionally making their profits vulnerable to the actions of others may seem inconsistent with our discussion on "make-or-buy" decisions. In that early chapter we argued that firms often forgo the advantages of buying inputs in the marketplace by making them in-house to protect their profits on their investment against exploitation by others. The difference in the two cases is important. When firms put their profits at risk as a hostage to consumers, those consumers cannot capture the profits for themselves. They can only destroy them, and their only motivation for doing so would be that the firm is no longer satisfying their demands. In the case where a firm incurs the disadvantage of producing in-house to protect its profits, the problem is that suppliers can actually capture those profits for themselves by acting opportunistically, or dishonestly. So in some cases protecting profits promotes honest dealing, and in other cases putting those profits at risk promotes honest dealing.

The importance business people attach to committing themselves to honesty sometimes leads them to put their profits in a position to be competed away by other firms that will benefit from doing so. Consider a situation where a firm has a patent on a high quality product that consumers would like to purchase at the advertised price, but a product that would be difficult to stop using because its use requires costly commitments. The fear of the potential buyers is that the seller will exploit the long-term patent

---

[5] Technically speaking, the "extra profit" we have in mind is dubbed "quasirents" by economists, and quasirents are the returns that can be made off a fixed investment over and above what can be earned elsewhere. These profits, or quasirents, can be extracted by opportunistic behavior because the investment's value is lower in some other activity. We use the term "profit" here and elsewhere because it is more familiar to general business readers and because the terms "rent" (or "quasirents") might be confused with the monthly payments businesses make for the use of their buildings.

[6] A number of years ago, one of the major pantyhose companies hired the famous football player, Joe Namath, to advertise their pantyhose by claiming that they were his favorite brand. This was surely not done to convince the public that Joe Namath actually wore a particular brand of pantyhose, or any pantyhose for that matter. A more plausible explanation is that the company wanted an advertisement that would get the public's attention and let people know that they were making enough money in the pantyhose business to hire Joe Namath, who was a very expensive spokesman at the time.

monopoly on the product by raising the price after the buyer commits to it at the attractive initial price. The seller may promise not to raise the price, but the buyer will be taking an expensive risk to trust the honesty of the promise. A long-term contract is possible, but it is difficult to specify all the contingencies under which a price increase (or decrease) would be justified. Also, such a contract can reduce the flexibility of the buyer as well as the seller, and legal action to enforce the contract is expensive.

Another possibility is for the seller to give up his or her monopoly position by licensing another firm to sell the product. By doing so the seller makes his or her promise to charge a reasonable price in the future credible, since if the seller breaks the promise the buyer can turn to an alternative seller. Giving up a monopoly position is a costly move of course, but it is exactly what semiconductor firms that have developed patented chips have done. To make credible their promise of a reliable and competitively priced supply of a new proprietary chip (the use of which requires costly commitments by the user), semiconductor firms have licensed such chips to competitive firms. Such a licensing arrangement is another example of making profits by way of a hostage intended to encourage honesty.[7]

The more difficult it is for consumers to determine the quality of a product or service, the more advantage there is in committing to honesty with hostage arrangements. Consider the case of repair work. When someone purchases repair work on their car, for example, they can generally tell if the work eliminates the problem. The car is running again, the rattle is gone, the front wheels now turn in the same direction as the steering wheel, etc. But few people know if the repair shop charged them for only the repairs necessary, or if it charged them for lots of parts and hours of labor when tightening a screw was all that was done. One way repair shops can reduce the payoff to dishonest repair charges is through joint ownership with the dealership selling the cars being repaired. In this way the owner of the dealership makes future car sales a hostage to honest repair work. Dealerships depend on repeat sales from satisfied customers, and an important factor in how satisfied people are with their cars is the cost of upkeep and repairs. The gains a dealership could realize from overcharging for repair work would be quickly offset by reductions in both repair business and car sales.

Automobiles are not the only products in which it is common to find repairs and sales tied together in ways that provide incentives for honest dealing. Many products come with guarantees entitling the buyer to repairs and replacement of defective parts for a specified period of time. These guarantees also serve as hostages against poor quality and high repair costs. Of course, guarantees not only provide assurance of quality, they provide protection against the failure of that assurance. Sellers often offer extra assurance, and the opportunity to reduce their risk, by selling a warranty with their product that extends the time, and often the coverage, of the standard guarantee.

---

[7]When Intel developed its 286 microprocessor in the late 1970s, it gave up its monopoly by licensing other firms to produce it [as discussed by Adam M. Brandenburger and Barry J. Nalebuff, Co-opetition (New York: Currency/Doubleday, 1996), pp. 105-106].

Moral Hazard and Adverse Selection

While guarantees and warranties reduce the incentive of sellers to act dishonestly, they create opportunities for buyers to benefit from less than totally honest behavior. These opportunities are present to one degree or another in all forms of insurance and come as two separate problems, one known as **moral hazard** (or the tendency of behavior to change after contracts are signed, resulting in unfavorable outcomes from the use of a good or service) and the other known as **adverse selection** (or the tendency of people to buy good or service when they know their characteristics are undesirable to sellers). Consider first the problem of moral hazard.

Knowing that a product is under guarantee or warranty can tempt buyers to use the product improperly and carelessly, and then blame the seller for the consequences. With this moral hazard in mind, sellers put restrictions on guarantees and warranties that leave buyers responsible for problems they are in the best position to prevent. For example, refrigerator manufacturers ensure against defects in the motor but not against damage to the shelves or finish. Similarly, automobile manufacturers ensure against problems in the engine and drive train (if the car has been properly serviced) but not against damage to the body and the seat covers. While such restrictions obviously serve the interests of sellers, they also serve the interests of buyers. When a buyer takes advantage of a guarantee by misrepresenting the cause of a difficulty with a product, all consumers pay because of higher costs to the seller. Buyers are in a prisoners' dilemma in which they are better off collectively using the product with care and not exploiting a guarantee for problems they could have avoided. But without restrictions on the guarantee each individual is tempted to shift the cost of their careless behavior to others.

Adverse selection is a problem associated with distortions arising from the fact that buyers and sellers often have different information that is relevant to a transaction. Most of this chapter has been concerned with the ways sellers commit themselves to honestly revealing the quality of products when they have more information about that quality than do buyers. But in the case of warranties it is the buyer who has crucial information that is difficult for the seller to obtain. Some buyers are harder on the product than average and others are easier on the product than average. The use of automobiles is the most obvious example. Some people drive in ways that greatly increase the probability that their cars will need expensive repair work, while others drive in ways that reduce that probability. If a car manufacturer offers a warranty at a price equal to the average cost of repairs, only those who know that their driving causes greater than average repair costs will purchase the warranty, which is therefore being sold at a loss. If the car manufacturer attempts to increase the price of the warranty to cover the higher than expected repair costs, then more people will drop out of the market leaving only the worst drivers buying the warranty.[8]

Even though people would like to be able to reduce their risks by purchasing warranties at prices that accurately reflect their expected repair bills, the market for these

[8]This warranty problem is similar to the lemon problem discussed earlier in this chapter, but in this case it is the buyers who are supplying the lemons in the form of their behavior.

warranties can obviously collapse unless sellers can somehow obtain information on the driving behavior of different drivers. If all buyers were honest in revealing this information they would be better off collectively. But because individual buyers have a strong motivation to claim they are easier on their cars than they actually are, sellers of warranties try to find indirect ways of securing honest information on the driving behavior of customers. For example, warranties on "muscle" cars that appeal to young males are either more expensive, or provide less coverage, than warranties on station wagons.

This section has focused primarily on business arrangements that motivate firms to deal honestly with customers, and our discussion of these arrangements is far from exhaustive. Honesty is also important in the interaction between shareholders and managers, employers and workers, and creditors and debtors, and many different types of arrangements exist that motivate trustworthy behavior in these relationships. Such business arrangements serve a variety of purposes such as marketing products, financing capital investment, and securing productive workers, but understanding any of them requires recognizing the importance business people attach to being able to commit themselves credibly to honesty in their dealings with others.

Concluding Comments

As we have argued, a market economy will overproduce goods and services that impose external costs on society. It will underproduce goods and services that confer external benefits. Sometimes, but not always, government intervention can be justified to correct for externalities. To be worthwhile, the benefits of action must outweigh the costs.

Some ways of dealing with external costs and benefits are more efficient than others. Even when government intervention in the market is clearly warranted, the method of intervening must be carefully selected.

Some critics of markets suggest that markets are bound to fail because of the gains to business from being dishonest, which implies a form of "externality." While we would be the first to recognize the pervasiveness of dishonest behavior, we also hasten to stress that markets have built-in incentives for people to be more honest that they might otherwise be.

Review Questions

1. The existence of external costs is not in itself a sufficient reason for government intervention in the production of steel. Why not?
2. "Population growth will lead to increased government control over people's behavior." Do you agree or disagree? Explain.
4. Developers frequently buy land and hold it on speculation; in effect they "bank" land. Should firms be permitted to buy and bank pollution rights in the same say? Would such a practice contribute to overall economic efficiency?

5. "If allowing firms to trade pollution rights lowers the cost of meeting pollution standards, it should also allow government to tighten standards without increasing costs." Do you agree or disagree? Why?
6. If businesses are permitted to sell pollution rights, should brokers in pollution rights be expected to emerge? Why or why not? Would such agents increase the efficiency with which pollution is cleaned up?
7. If pollution rights are traded, should the government impose a price ceiling on them? Would such a system contribute to the efficient allocation of resources?
8. If you were a producer, which method of pollution control would you favor, the setting of government standards or the auction of pollution rights by government? Why?

CHAPTER 8

Consumer Choice and Demand in Traditional and Network Markets

It is not the province of economics to determine the value of life in "hedonic units" or any other units, but to work out, on the basis of the general principles of conduct and the fundamental facts of social situation, the laws which determine prices of commodities and the direction of the social economic process. It is therefore not quantities, not even intensities, of satisfaction with which we are concerned. . . .or any other absolute magnitude whatever, but the purely relative judgment of comparative significance of alternatives open to choice.

Frank Knight

People adjust to changes in some economic conditions with a reasonable degree of predictability. When department stores announce lower prices, customers will pour through the doors. The lower the prices go, the larger the crowd will be. When the price of gasoline goes up, drivers will make fewer and shorter trips. If the price stays up, drivers will buy smaller, more economical cars. Even the Defense Department will reduce its planned purchases when prices rise.

Behavior that is not measured in dollars and cents is also predictable in some respects. Students who stray from the sidewalks to dirt paths on sunny days stick to concrete when the weather is damp. Professors who raise their course requirements and grading standards find their classes are shrinking in size. Small children shy away from doing things for which they have recently been punished. When lines for movie tickets become long, some people go elsewhere for entertainment.

On an intuitive level you find these examples reasonable. Going one step beyond intuition, the economist would say that such responses are the predictable consequences of rational behavior. That is, people who desire to maximize their utility can be expected to respond in these ways. Their responses are governed by the law of demand, a concept we first introduced in Chapter 3 and now take up in greater detail.

Predicting Consumer Demand

The assumptions about rational behavior described early in the book provide a good general basis for explaining behavior. People will do those things whose expected benefits exceed their expected costs. They will avoid doing things for which the opposite is true. By themselves, however, such assumptions do not allow us to predict future

behavior. The law of demand, which is a logical consequence of the assumption of rational behavior, does allow us to make predictions.

The alert reader may sense an inconsistency in logic. Rational behavior is based on the existence of choice, but a true choice must be free—it cannot be predetermined or predicted. If we can predict a person's behavior, can that individual be free to choose?

Choice is not completely free, nor is complete freedom required by the concept of rationality. As discussed earlier, the individual's choices are constrained by time and by physical and social factors that restrict his or her opportunities. There are limits to a person's range of choice. Freedom exists within those limits.

Our ability to predict is also limited. We cannot specify with precision every choice the individual will make. For instance, we cannot say anything about what Judy Schwartz wants or how much she wants the things she does. Before we can employ the law of demand, we must be told what she wants. Even given that knowledge, we can only indicate the general direction of her behavior. Theory does not allow us to determine how fast or how much her behavior will change.

To see how consumer behavior can be predicted, we will derive the law of demand from the behavior of an individual consumer.

Rational Consumption: The Concept of Marginal Utility

The essence of the economist's notion of rational behavior can be summed up this way: more goods and services are preferable to less (assuming that the goods and services are desired). This statement implies that the individual will use his entire income, in consumption or in saving or in some combination of the two, to maximize his satisfaction. It also implies that the individual will use some method of comparing the value of various goods.

Generally speaking, the value the individual places on any one unit of a good depends on the number of units already consumed. For example, you may be planning to consume two hot dogs and two Cokes for your next meal. Although you may pay the same price for each unit of both goods, there is no reason to assume that you will place the same value on each. The value of the second hot dog—its marginal utility—will depend on the fact that you have already eaten one. The formula for marginal utility is

$$MU = \frac{\textit{change in total utility}}{\textit{change in quantity consumed}}$$

Achieving Consumer Equilibrium

Marginal utility determines the variety of a quantity of goods and services you consume. The rule is simple. If the two goods, Cokes and hot dogs, both have the same price, you will allocate your income so that the marginal utility of the last unit of each will be equal. Mathematically, the formula can be stated as

$$MU_c = MU_h$$

Where MU_c equals the marginal utility of a Coke and MU_h equals the marginal utility of a hot dog.

If you are rational, and if the price of a Coke is the same as the price of a hot dog, the last Coke you drink will give you the same amount of enjoyment as the last hot dog you eat. When the marginal utilities of goods purchased by the consumer are equal, the resulting state is called consumer equilibrium. **Consumer equilibrium** is a state of stability in consumer purchasing patterns in which the individual has maximized his or her utility. Unless conditions—income, taste, or prices—change, the consumer's buying patterns will tend to remain the same.

An example will illustrate how equilibrium is reached. Suppose for the sake of simplicity that you can buy only two goods, Cokes and hot dogs. Suppose further that one of each cost the same price, $1, and you are going to spend your whole income. (How much your total income is and how many units of Coke or hot dogs you will purchase is unimportant. We simply assume that you purchase some combination of those two goods.) We will also assume that utility (joy, satisfaction) can be measured. As you remember from an earlier chapter, a unit of satisfaction is called a util. Finally, suppose that the marginal utility of the last Coke you consume is equal to 20 utils, and the marginal utility of the hot dog is 10 utils. Obviously you have not maximized your utility, for the marginal utility of your last Coke is greater than (>) the marginal utility of your last hot dog:

$$MU_c > MU_h$$

You could have purchased one less hot dog and used the dollar saved the to buy an additional Coke. In doing so, you would have given up 10 utils of satisfaction (the marginal utility of the last hot dog purchased), but you would have acquired an additional 20 utils from the new Coke. On balance, your total utility would have risen by 10 utils (20 – 10). If you are rational, you will continue to adjust your purchases of Coke and hot dogs until their marginal utilities are equal.

Even if you would prefer to spend your first dollar on a hot dog, after eating several you might wish to spend your next dollar on a Coke. Purchases can be adjusted until they reach equilibrium because as more of a good is purchased, its relative marginal utility decreases—a phenomenon known as the law of diminishing marginal utility. According to the **law of diminishing marginal utility**, as more of a good is consumed, its marginal utility or value relative to the marginal value of the good or goods given up eventually diminishes. Thus, if $MU_h > MU_c$, and MU_h falls relative to MU_c as more hot dogs and fewer Cokes are consumed, sooner or later the result will be $MU_h = MU_c$.

Adjusting for Differences in Price and Unit Size

Cokes and hot dogs are not usually sold at exactly the same price. To that extent, our analysis has been unrealistic. If we drop the assumption of equal prices, the formula for maximization of utility becomes:

$$\frac{MU_c}{P_c} = \frac{MU_h}{P_h}$$

Where MU_c equals the marginal utility of a Coke, MU_h the marginal utility of a hot dog, P_c the price of a Coke, a P_h the price of a hot dog. This is the same formula we used before, but because the price of the goods was the same in that example, the denominators canceled out. When prices differ, the denominator must be retained. The consumer must allocate his or her money so that the last penny spent on each commodity yields the same amount of satisfaction.

Suppose a Coke costs $0.50 and the price of a hot dog is $1. If you buy hot dogs and Cokes for lunch and the marginal utility of the last Coke and hot dog you consume are the same, say 15 utils, you will not be maximizing your satisfaction. In relation to price, you will value your Coke more than your hot dog. That is, MU_c/P_c (or 15 utils/$0.50) exceeds MU_h/P_h (or 15 utils/$1). You can improve your welfare by eating fewer hot dogs and drinking more Cokes. By giving up a hot dog, you can save a dollar, which you can use to buy two Cokes. You will lose 15 utils by giving up the hot dog, something you would probably prefer not to do. You will regain that loss with the next Coke purchased, however, and the one after that will permit you to go beyond your previous level of satisfaction.

Therefore, if you are rational, you will adjust your purchases until the utility-price ratios of the two goods are equal. As you consume more Coke, the relative value of each additional Coke will diminish. If you reach a point where the next Coke gives you 10 utils and the next hot dog yields 20 utils, you will no longer be able to increase your satisfaction by readjusting your purchases. By giving up the next hot dog, you save $1 and lose 20 utils of satisfaction. Now the most you can accomplish by using that $1 to buy two Coke instead is to recoup your loss of 20 utils. In fact, the value of the second new Coke may be less than 10 utils, so you may actually lose by giving up the hot dog.

So far we have been talking in terms of buying whole units of Cokes and hot dogs, but the same principles apply to other kinds of choices as well. Marginal utility is involved when a consumer chooses a 12-ounce rather than a 16-ounce can of Coke, or a regular-size hot dog rather than a foot-long hot dog. The concept could also be applied to the decision whether to add cole slaw and chili to the hot dog. The pivotal question the consumer faces in all these situations is whether the marginal utility of the additional quantity consumed is greater or less than the marginal utility of other goods that can be purchased for the same price.

Most consumers do not think in terms of utils when they are buying their lunch, but in a casual way, they do weigh the alternatives. Suppose you walk into a snack bar. If your income is unlimited, you have no problem. If you can only spend $3 for lunch, however, your first reaction may be to look at the menu and weigh the marginal values of the various things you can eat. If you have twenty cents to spare, do you not find yourself mentally asking whether the difference between a large Coke and a small one is worth more to you than lettuce and tomato on your hamburger? (If not, why do you choose a small Coke instead of a large one?) You are probably so accustomed to making decisions of this sort that you are almost unaware of the act of weighing the marginal values of the alternatives.

Consumers do not usually make choices with conscious precision. Nor can they achieve a perfect equilibrium—the prices, unit sizes, and values of the various products available may not permit it. They are trying to come as close to equality as possible, The economist's assumption is that the individual will move toward equality, not that he will always achieve it.

Changes in Price and the Law of Demand

Suppose your marginal utility for Coke and hot dogs is as shown in the table below.

Unit Consumed	Marginal Utility of Cokes (at $0.50)	Marginal Utility of Hot Dogs (at $1)
First	10 utils	30 utils
Second	9 utils	15 utils
Third	3 utils	12 utils

If a Coke is priced at $0.50 and a hot dog at $1, $3 will buy you two hot dogs and two Cokes—the best you can do with $3 at those prices. Now suppose the price of Coke rises to $0.75 and the price of hot dogs falls to $0.75. With a budget of $3 you can still buy two hot dogs and two Cokes, but you will no longer be maximizing your utility. Instead you will be inclined to reduce your consumption of Coke and increase your consumption of hot dogs.

At the old prices, the original combination (two Cokes and two hot dogs) gave you a total utility of only 64 utils (45 from hot dogs and 19 from Coke). If you cut back to one Coke and three hot dogs now, your total utility will rise to 67 utils (57 from hot dogs and 10 from Coke). Your new utility-maximizing combination—the one that best satisfies your preferences—will therefore be one Coke and three hot dogs. No other combination of Coke and hot dogs will give you greater satisfaction. (Try to find one.)

To sum up, if the price of hot dogs goes down relative to the price of Coke, the rational person will buy more hot dogs. If the price of Coke rises relative to the price of hot dogs, the rational person will buy less Coke. This principle will hold true for any good or service and is commonly known as the law of demand. The **law of demand** states the assumed inverse relationship between product price and quantity demanded, everything else held constant. If the relative price of a good falls, the individual will buy more of the good. If the relative price rises, the individual will buy less.

Figure 8.1 shows the demand curve for Coke—that is, the quantity of Coke purchased at different prices. The inverse relationship between price and quantity is reflected in the curve's downward slope. If the price falls from $1 to $0.75, the quantity the consumer will buy increases from two Cokes to three. The opposite will occur if the price goes up.

Thus the assumption of rational behavior, coupled with the consumer's willingness and ability to substitute less costly goods when prices go up, leads to the law of demand.

We cannot say how many Cokes and hot dogs a particular person will buy to maximize his or her satisfaction. That depends on the individual's income and preferences, which depend in turn on other factors (how much he likes hot dogs, whether he is on a diet, and how much he worries about the nutritional deficiencies of such a lunch). We can predict the general response, whether positive or negative, to a change in prices.

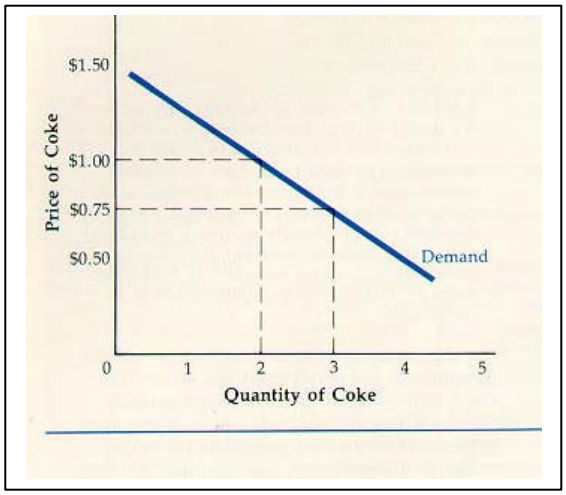


FIGURE 8.1 The Law of Demand

Price varies inversely with the quantity consumed, producing a downward-sloping curve like this one. If the price of Coke falls from $1 to $0.75, the consumer will buy three Cokes instead of two.

Price is whatever a person must give up in exchange for a unit of goods or services purchased, obtained, or consumed. It is a rate of exchange and is typically expressed in dollars per unit. Note that price is not necessarily the same as cost. In an exchange between two people—a buyer and a seller—the price at which a good sells can be above or below the cost of producing the good. What the buyer gives up to obtain the good does not have to match what the seller-producer gives up in order to provide the good.

Nor is price always stated in dollars and cents. Some people have a desire to watch sunsets—a want characterized by the same downward-sloping demand curve as the one for Coke. The price of the sunset experience is not money. Instead it may be the lost opportunity to do something else, or the added cost and trouble of finding a home that will offer a view of the sunset. (In that case, price and cost are the same because the buyer and the producer are one and the same.) The law of demand will apply nevertheless. The individual will spend some optimum number of minutes per day watching the sunset and will vary that number of minutes inversely with the price of watching.

From Individual Demand to Market Demand

Thus far we have discussed demand solely in terms of the individual's behavior. The concept is most useful, however, when applied to whole markets or segments of the population. **Market demand** is the summation of the quantities demanded by all consumers of a good or service at each and every price during some specified time period. To obtain the market demand for a product, we need to find some way of adding up the wants of the individuals who collectively make up the market.

The market demand can be shown graphically as the horizontal summation of the quantity of a product each individual will buy at each price. Assume that the market for Coke is composed of two individuals, Anna and Betty, who differ in their demand for Coke, as shown in Figure 8.2. The demand of Anna is D_A and the demand of Betty is D_B. Then to determine the number of Cokes both of them will demand at any price, we simply add together the quantities each will purchase, at each price (see Table 8.1). At a price of $11, neither person is willing to buy any Coke; consequently, the market demand must begin below $11. At $9, Anna is still unwilling to buy any Coke, but Betty will buy two units. The market quantity demanded is therefore two. If the price falls to $5, Anna wants two Cokes and Betty, given her greater demand, wants much more, six. The two quantities combined equal eight. If we continue to drop the price and add the quantities bought at each new price, we will obtain a series of market quantities demanded. When plotted on a graph they will yield curve D_{A+B} , the market demand for Coke (see Figure 8.2).

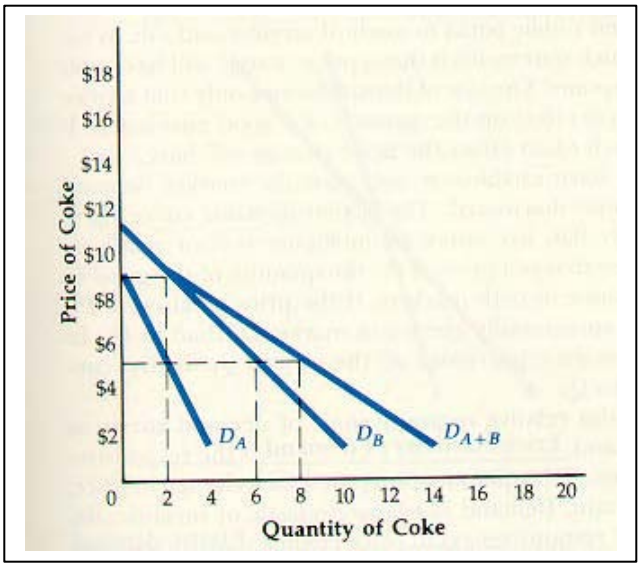


FIGURE 8.2 Market Demand Curve

The market demand curve for Coke, D_{A+B}, is obtained by summing the quantities that individuals A and B are willing to buy at each and every price (shown by the individual demand curves D_A and D_B).

This is, of course, an extremely simple example, since only two individuals are involved. The market demand curves for much larger groups of people, however, are derived in essentially the same way. The demands of Fred, Marsha, Roberta, and others would be added to those of Anna and Betty. As more people demand more Coke, the market demand curve flattens out and extends further to the right.

Elasticity: Consumers' Responsiveness to Price Changes

In the media and in general conversation, we often hear claims that a price change will have no effect on purchases. Someone may predict that an increase in the price of prescription drugs will not affect people's use of them. The same remark is heard in connection with many other goods and services, from gasoline and public parks to medical services and salt. What people usually mean by such statements is that a price

change will have only a slight effect on consumption. The law of demand states only that a price change will have an inverse effect on the quantity of a good purchased. It does not specify how much of an effect the price change will have.

TABLE 8.1 Market Demand for Coke

Price of Coke (1)	Quantity Demanded by Anna (D_A) (2)	Quantity Demanded by Betty (D_B) (3)	Quantity Demanded by both Anna and Betty (D_{A+B}) (4)
$11	0	0	0
10	0	1	1.0
9	0	2	2.0
8	0.5	3	3.5
7	1.0	4	5.0
6	1.5	5	6.5
5	2.0	6	8.0
4	2.5	7	9.5
3	3.0	8	11.0
2	3.5	9	12.5
1	4.0	10	14.0

Note: The market demand curve, D_{A+B}, in Figure 8.2 is obtained by plotting the quantities in column (4) against their respective prices in column (1).

In other words, we have established only that the market demand curve for a good will slope downward. The actual demand curve for a product may be relatively flat, like curve D_1 in Figure 8.3, or relatively steep, like curve D_2. Notice that at a price of P_1, the quantity of the good or service consumed is the same in both markets. If the price is raised to P_2, however, the response is substantially greater in market D_1 than in D_2. In D_1, consumers will reduce their purchases all the way to Q_1. In D_2, consumption will drop only to Q_2.

Economists refer to this relative responsiveness of demand curves as the price elasticity of demand. **Price elasticity of demand** is the responsiveness of consumers, in terms of the quantity purchased, to a change in price, everything else held constant. Demand is relatively elastic or inelastic, depending on the degree responsiveness to price change. **Elastic demand** is a relatively sensitive consumer response to price changes. If the price goes up or down, consumers will respond with a strong decrease or increase in the quantity demanded. Demand curve D_1 in Figure 8.3 may be characterized as relatively elastic. **Inelastic demand** is a relatively insensitive consumer response to price changes. If the price goes up or down, consumers will respond with only a slight decrease or increase in the quantity demanded. Demand curve D_2 in Figure 8.3 is relatively inelastic.

The elasticity of demand is a useful concept, but our definition is imprecise. What do we mean by "relatively sensitive" or "relatively insensitive"? Under what

circumstances is consumer response sensitive or insensitive? There are two ways to add precision to our definition. One is to calculate the effect of a change in price on total consumer expenditures (which must equal producer revenues). The other is to develop mathematically values for various levels of elasticity. We will deal with each in turn.

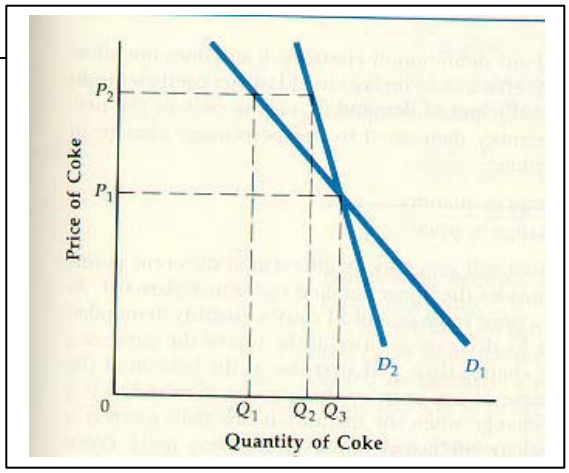


FIGURE 8.3 Elastic and Inelastic Demand

Demand curves differ in their relative elasticity. Curve D_1 is more elastic than curve D_2, in the sense that consumers on curve D_1 are more responsive to a price change than are consumers on curve D_2.

Analyzing Total Consumer Expenditures

An increase in the price of a particular product can cause consumers to buy less. Whether total consumer expenditures rise, fall, or stay the same, however, depends on the extent of the consumer response. Many people assume that businesses will charge the highest price possible to maximize profits. Although they sometimes do, high prices are not always the best policy. For example, if a firm sells fifty units of a product for $1, its total revenue (consumers' total expenditures) for the product will be $50 (50 x $1). If it raises the price to $1.50 and consumers cut back to forty units, its total revenue could rise to $60 (40 x $1.50). If consumers are highly sensitive to price changes for this particular good, however, the fifty-cent increase may lower the quantity sold to thirty units. In that case total consumer expenditures would fall to $45 ($1.50 x 30).[1]

[1] To prove this result, let's look at marginal revenue MR, or the change in total revenue in response to a change in quantity Q. Taking the derivative of $P(Q) \cdot Q$ with respect to Q, we obtain

$$\text{MR} = \frac{d[P(Q) \bullet Q]}{dQ} = P(Q) + \frac{dP}{dQ} \bullet Q$$

Factoring price out of the right-hand side of this equation gives us

$$\text{MR} = P\left[1 + \frac{dP}{dQ} \bullet \frac{Q}{P}\right]$$

which, because $E = -\left(\frac{dQ}{dP}\right)$, is the same as

The opposite can also happen. If a firm establishes a price of $1.50 and then lowers it to $1, the quantity sold may rise, but the change in total consumer expenditures will depend on the degree of consumer response. In other words, consumer responsiveness determines whether a firm should raise or lower its price. (We will return to this point later.)

We can define a simple rule of thumb for using total consumer expenditures to analyze the elasticity of demand. Demand is *elastic*:

- if total consumer expenditures rise when the price falls, or
- if total consumer expenditures fall when the price rises.

Demand is *inelastic*:

- if total consumer expenditures rise when the price rises, or
- if total consumer expenditures fall when the price falls.

Determining Elasticity Coefficients

Although we have refined our definition of elasticity, it still does not allow us to distinguish degrees of elasticity or inelasticity. Elasticity coefficients do just that. The **elasticity coefficient of demand** (E_d) is the ratio of the percentage change in the quantity demanded to the percentage change in price. Expressed as a formula,

$$E_d = \frac{\text{percentage change in quantity}}{\text{percentage change in price}}$$

The elasticity coefficient will generally be different a different points on the demand curve. Consider the linear demand curve in Figure 8.4. At every point on the curve, a price reduction of $1 causes quantity demanded by rise by ten units, but a $1 decrease in price at the top of the curve is a much smaller percentage change than a $1 decrease at the bottom of the curve. Similarly, an increase of ten units in the quantity demanded is a much larger percentage change when the quantity is low than when it is high. Therefore the elasticity coefficient falls as consumers move down their demand curve. Generally, a straight-line demand curve has an inelastic range at the bottom, a unitary elastic point in the center, and an elastic range at the top.[2]

$$\text{MR} = P\left[1 - \frac{1}{E}\right] \begin{array}{l} > 0 \text{ if } E > 1 \\ = 0 \text{ if } E = 1 \\ < 0 \text{ if } E < 1 \end{array}$$

From this it follows immediately that an increase in Q (a decrease in P) increases total revenue if $E > 1$, has no effect on total revenue if $E = 1$, and reduces total revenue if $E < 1$.

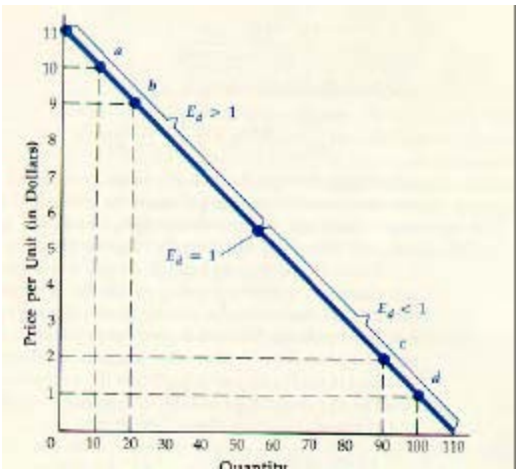
[2] To prove this, we recognize that the equation for a linear domain curve can be expressed mathematically as

$$P = \text{A} - \text{B}Q$$

where P represents price, Q is quantity demanded, and A and B are positive constants. The total revenue associated with this demand curve is given by

FIGURE 8.4 Changes in the Elasticity Coefficient

The elasticity coefficient decreases as a firm moves down the demand curve. The upper half of a linear demand curve is elastic, meaning that the elasticity coefficient is greater than one. The lower half is inelastic, meaning that the elasticity coefficient is less than one. This means that the middle of the linear demand curve has an elasticity coefficient equal to one.



There are two formulas for elasticity, one for use at specific points on the curve and one for measuring average elasticity between two points, called arc elasticity. The formula for point elasticity, which is used for very small changes in price, is:

$$E_d = \frac{\text{change in quantity demanded}}{\text{initial quantity demanded}} \div \frac{\text{change in price}}{\text{initial price}}$$

or

$$E_d = \frac{Q_1 - Q_2}{Q_1} \div \frac{P_1 - P_2}{P_1}$$

The formula for arc elasticity is:

$$PQ = AQ - BQ^2$$

The marginal revenue is obtained by taking the derivative of total revenue with respect to Q or

$$MR = A - 2BQ$$

From footnote 1, we know that when marginal revenue is equal to 0, elasticity is equal to 1. From Equation (2) here, this implies that $E = 1$ when

$$A - 2BQ = 0$$

or when

$$Q = \frac{1}{2} \bullet \frac{A}{B}$$

From Equation (1) we know that when the demand curve intersects the Q axis, $P = 0$ and

$$Q = \frac{A}{B}$$

Thus, with a linear demand curve, $E = 1$ when Q is one-half the distance between $Q = 0$ and the Q that drives price down to 0. The reader is invited to prove that $E > 1$ when $Q < ½ \bullet A/B$, and that $E < 1$ when $Q > ½ \bullet A/B$.

$$E_d = \frac{Q_1 \text{ -- } Q_2}{\frac{1}{2}(Q_1 + Q_2)} \div \frac{P_1 - P_2}{\frac{1}{2}(P_1 \text{ -- } P_2)}$$

Where the subscripts 1 and 2 represent two distinct points, or prices, on the demand curve. Note that although the calculated elasticity is always negative, economists, by convention, speak of it as a positive number. Economists, in effect, use the absolute value of elasticity.

The change can be illustrated by computing the arc elasticity between two sets of points, *ab* and *cd*. Arc elasticity between points *a* and because:

$$E_d = \frac{10 - 20}{\frac{1}{2}(10 + 20)} \div \frac{10 - 9}{\frac{1}{2}(10 + 9)} = -\frac{10}{15} \div \frac{1}{9.5} = -\frac{95}{15} = -6.33$$

or 6.33 in absolute value.

Arc elasticity between points *c* and *d*:

$$E_d = \frac{90 - 100}{\frac{1}{2}(90 + 100)} \div \frac{2 - 1}{\frac{1}{2}(2 + 1)} = -95 \div 1.5 = -0.16 \text{ or } 0.16 \text{ in absolute value.}$$

Elasticity coefficients can tell us much at a glance. When the percentage change in quantity is greater than the percentage change in price, an elasticity coefficient that is greater than 1.0 results. In these cases, demand is said to be elastic. When the percentage change in quantity is less than the percentage change in price, the elasticity coefficient will be less than 1.0. Demand is said to be inelastic. When the percentage change in the price is equal to the percentage change in quantity, the elasticity coefficient is 1.0, and demand is unitary elastic.[3] In short:

Elastic demand:	$E_d > 1$
Inelastic demand:	$E_d < 1$
Unitary elastic demand:	$E_d = 1$

Elasticity coefficients enable economists to make accurate comparisons. A demand with an elasticity coefficient of 1.75 is more elastic than one with an elasticity coefficient of 1.55. A demand with a coefficient of 0.25 is more inelastic than one with a coefficient of 0.78.

Although elasticity coefficients are useful for some purposes, their accuracy depends on data that are often less than precise. In the real world, there is constant change in the nonprice variables that influence how much of any product consumers want. It is extremely difficult for economists to separate the effects of a change in price

[3] Remember that all elasticity coefficients are negative and are preceded by a minus sign. (The demand curve has a negative slope.) economists generally omit the minus sign, as we have seen.

from all the other forces operating in the marketplace. Small differences in elasticity coefficients may reflect the imperfections of statistical analysis rather than true differences in consumer responsiveness to price.

Elasticity, Not the Same as Slope

Students often confuse the concept of elasticity of demand with the slope of the demand curve. A comparison of their mathematical formulas, however, shows they are quite different.

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{\text{change in price}}{\text{change in quantity}}$$

$$\text{elasticity} = \frac{\text{percentage change in quantity}}{\text{percentage change in price}}$$

The confusion is understandable. The slope of a demand curve does say something about consumer's responsiveness: it shows how much the quantity consumed goes up when the price goes down by a given amount. Slope is an unreliable indicator of consumer responsiveness, however, because it varies with the units of measurement for price and quantity. For example, suppose that when the price rises from $10 to $20, quantity demanded decreases from 100 to 60. The slope is –1/4.

$$\text{slope} = \frac{-10}{40} = \frac{-1}{4}$$

If a price is measured in pennies instead of dollars, however, the slope comes out to –25.

$$\text{slope} = \frac{-1000}{40} = \frac{-25}{1}$$

No matter how the price is measured, the arc elasticity of demand remains –0.75. Furthermore, two parallel demand curves of identical slope will not have the same elasticity coefficients. For example, consider the two curves in Figure 8.5. When the price falls from $5 to $4, the quantity demanded rises by the same amount for each curve: ten units. Yet the percentage change in quantity is substantially lower for D_2 than for D_1. (A rise from seventy to eighty is not nearly as dramatic in percentage terms as a rise from twenty-five to thirty-five.) Thus the elasticity coefficient is lower for demand curve D_2.

Be careful not to judge the elasticity of demand by looking at a curve's slope.

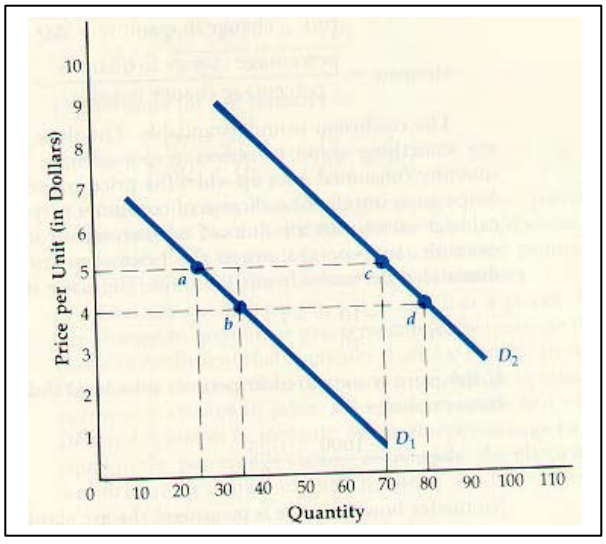
Applications of the Concept of Elasticity

Elasticity of demand is particularly important to producers. Together with the cost of production, it determines the prices firms can charge for their products. We have seen that an increase or decrease in price can cause total consumer expenditures to rise, fall, or remain the same, depending on the elasticity of demand. Thus if a firm lowers its price and incurs greater production costs (because it is producing and selling more units), it may still increase its profits. As long as the demand curve is elastic, revenues can (but will not necessarily) go up more than costs. Over the last three decades, the American Telephone and Telegraph Company has frequently lowered its prices on long-distance calls. To justify those decisions, AT&T had to reason that demand was sufficiently elastic to produce revenues that would more than cover the cost of servicing the extra calls. During the 1950s a 1960s, many electric power companies requested rate reductions for the same reason.

Producers of concerts and dances estimate the elasticity of demand when they establish the price of admission. If admission costs $10, tickets may be left unsold. At a lower price, say $7, attendance and profits may be higher. Even if costs rise (for extra workers and more programs), revenues can still rise more.

FIGURE 8.5 Two Parallel Curves Do Not Have the Same Elasticity

Even two parallel demand curves of the same slope do not have the same elasticity. Although a given change in price—for example, a $1 change—will produce the same unit change in quantity demanded, the percentage change will differ. Here, a drop in price from $5 to $4 produces a ten-unit gain in quantity demanded on both curves D_1 and D_2. A ten-unit increase in sales represents a lower percentage change at an initial sales level of seventy (curve D_2).



The difference in the elasticity of the two curves can be illustrated by computing the arc elasticity between two sets of points, *ab* on curve D_1 and *cd* on curve D_2. Arc elasticity between points *a* and *b*:

$$E_d = \frac{25 - 35}{\frac{1}{2}(25 + 35)} \div \frac{5 - 4}{\frac{1}{2}(5 + 4)} = 1.50$$

Arc elasticity between points *c* and *d*:

$$E_d = \frac{70 - 80}{\frac{1}{2}(70 + 80)} \div \frac{5 - 4}{\frac{1}{2}(5 + 4)} = 0.60$$

Government too must consider elasticity of demand, for the consumer's demand for taxable items is not inexhaustible. If a government raises excise taxes on cars or jewelry too much, it may end up with lower tax revenues. The higher tax, added to the final price of the product, may cause a negative consumer response. It is no accident that

the heaviest excise taxes are usually imposed on goods for which the demand tends to be inelastic, such as cigarettes and liquor.

The same reasoning applies to property taxes. Many large cities have tended to underestimate the elasticity of demand for living space. Indeed, a major reason for the recent migration from city to suburbs in many metropolitan areas has been the desire of residents to escape rising tax rates. By moving just outside a city's boundaries, people can retain many of the benefits a city provides without actually paying for them. This movement of city dwellers to the suburbs lowers the demand for property within the city, undermining property values and destroying the city's tax base. Thus, if governments wish to maintain their tax revenues, they have to pay attention to the elasticity of demand for living in their jurisdictions.

Determinants of the Price Elasticity of Demand

So far our analysis of elasticity has presumed that consumers are able to respond to a price change. However, consumers' ability to respond can be affected by various factors, such as the number of substitutes and the amount of time consumers have to respond to a change in price by shifting to other products or producers.

Substitutes

Substitutes allow consumers to respond to a price increase by switching to another good. If the price of orange juice goes up, you are not required to go on buying it. You can substitute a variety of other drinks, including water, wine, and soda.

The elasticity of demand for any good depends very much on what substitutes are available. The existence of a large number and variety of substitutes means that demand is likely to be elastic. That is, if people can switch easily to another product that will yield approximately the same value, many will do so when faced with a price increase. The similarity of substitutes—how well they can satisfy the same basic want—also affects elasticity. The closer a substitute is to a product, the more elastic demand for the product will be. If there are no close substitutes, demand will tend to be inelastic. What we call necessities are often things that lack close substitutes.

Few goods have no substitutes at all. Because there are many substitutes for orange juice—soda, wine, prune juice, and so on—we would expect the demand for orange juice to be more elastic than the demand for salt, which has fewer viable alternatives. Yet even salt has synthetic substitutes. Furthermore, though human beings need a certain amount of salt to survive, most of us consume much more than the minimum and can easily cutback if the price of salt rises. The extra flavor that salt adds is a benefit that can be partially recouped by buying other things.

At the other extreme from goods with no substitutes are goods with perfect substitutes. Perfect substitutes exist for goods produced by an individual firm engaged in perfect competition. An individual wheat farmer, for example, is only one among thousands of producers of essentially the same product. The wheat produced by others is

a perfect substitute for the wheat produced by the single farmer. Perfect substitutability can lead to perfect elasticity of demand.

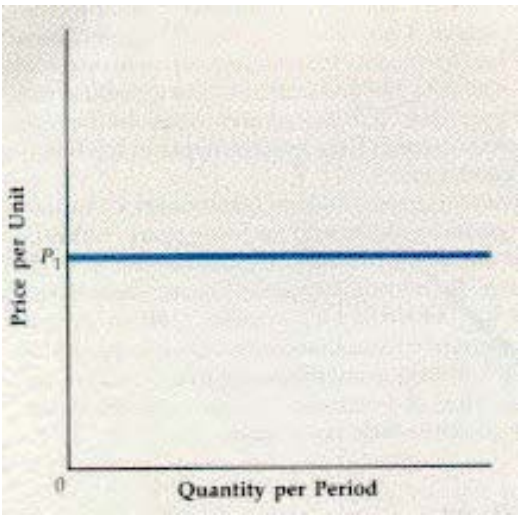
The demand curve facing the perfect competitor is horizontal, like the one in Figure 8.6. If the individual competitor raises his price even a minute percentage above the going market price, consumers will switch to other sellers. The elasticity coefficient of such a horizontal demand curve is infinite. Thus this demand curve is described as perfectly elastic. A **perfectly elastic demand** is a demand that has an elasticity coefficient of infinity. It is expressed graphically as a curve horizontal to the X-axis.

Time

Consumption requires time. Accordingly, a demand curve must describe some particular time period. Over a very short period of time—say a day—the demand for a good may not react immediately. It takes time to find substitutes. With enough time, however, consumers will respond to a price increase. Thus a demand curve that covers a long period will be more elastic than one for a short period.

FIGURE 8.6 Perfectly Elastic Demand

A firm that has many competitors may lose all its sales if it increases its price even slightly. Its customers can simply move to another producer. In that case its demand curve is horizontal, with an elasticity coefficient of infinity.



Oil provides a good example of how the elasticity of demand can change over time. In 1973 Arab oil producers raised the price of their crude oil, and domestic oil producers followed suit. For a time consumers were caught. Drivers were stuck with big, gas-guzzling cars and with suburban homes located far from their work places. Automakers were tooled up to produce big cars, not subcompacts. Over the long term, however, alternative modes of transportation became available and alternative sources of energy were found. People altered their lifestyles, walking or riding bicycles to work. The long-term demand curve for oil is much more elastic than the short-term demand curve.

Changes in Demand

The determinants of the elasticity of demand are fewer and easier to identify than the determinants of demand itself. As we saw in Chapter 3, the demand for almost all goods is affected in one way or another by (1) consumer incomes; (2) the prices of other goods; (3) the number of consumers; (4) expectations concerning future prices and incomes; and (5) that catchall variable, consumer tastes and preferences. Additional variables apply in differing degrees to different goods. The amount of ice cream and the number of golf balls bought both depend on the weather (very few golf balls are sold at the North Pole). The number of cribs demanded depends on the birthrate. Together all these variables determine the position of the demand curve. If any variable changes, so will the position of the demand curve.

We saw in Chapter 3 that if consumer preference for a product—say, blue jeans—increases, the change will be reflected in an outward movement of the demand curve (see Figure 8.7). That is what happened during the late 1960s, when college students' tastes changed and wearing faded blue jeans became chic. By definition, such a change in taste means that consumers are willing to buy more of the good at the going market price. If the price is P_1, the quantity demanded will increase from Q_2 to Q_3. A change in tastes can also mean that people are willing to buy more jeans at each and every price. At P_2 they are now willing to buy Q_2 instead of Q_1 blue jeans. We can infer from this pattern that consumers are willing to pay a higher price for any given quantity. In Figure 8.7, the increase in demand means that consumers are willing to pay as much as P_2 for Q_2 pairs of jeans, whereas formerly they would pay only P_1. (If consumers' tastes change in the opposite direction, the demand curve moves downward to the left, as in Figure 8.8, a quantity demanded at a given price decreases.)

Whether demand increases or decreases, the demand curve will still slope downward. Everything else held constant, people will buy more of the good at a lower price than a higher one. To assume that other variables will remain constant, of course, is unrealistic because markets are generally in a state of flux. In the real world, all variables just do not stay put to allow the price of a good to change by itself. Even if conditions change at the same time that price changes, the law of demand tells us that a decrease in price will lead people to buy more than they would otherwise, and an increase in price will lead them to buy less.

For example, in Figure 8.8, the demand for blue jeans has decreased, because consumers are less willing to buy the product. A price reduction can partially offset the decline in demand. If producers lower their price from P_2 to P_1, quantity demanded will fall only to Q_2 instead of Q_1. Although consumers are buying fewer jeans than they once did (Q_2 as opposed to Q_3) because of changing tastes, the law of demand still holds. Because of the price change, consumers have increased their consumption over what it would otherwise have been.

A change in consumer incomes will affect demand in more complicated ways. The demand for most goods, called normal goods, increases with income. A **normal good or service** is any good or service for which demand rises with an increase in income and falls with a decrease in income. The demand for a few luxury goods actually outstrips increases in income. A **luxury good or service** is any good or service for which

demand rises proportionally faster than income. An **inferior good or service** is any good or service for which demand falls with an increase in income and rises with a decrease in income. Beans are an example of a good many people would consider inferior. People who rely on beans as a staple or filler food when their incomes are low may substitute meat and other higher-priced foods when their incomes rise.

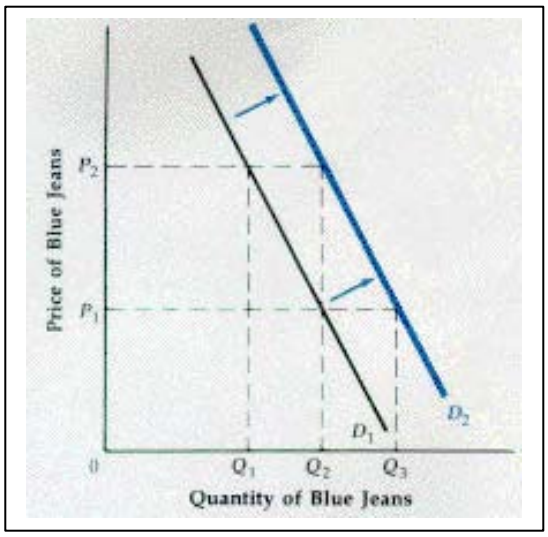


FIGURE 8.7 Increase in Demand

When consumer demand for blue jeans increases, the demand curve shifts from D_1 to D_2. Consumers are now willing to buy a larger quantity of jeans at the same price, or the same quantity at a higher price. At price P_1, for instance, they will buy Q_3 instead of Q_2. And they are now willing to pay P_2 for Q_2 jeans, whereas before they wold pay only P_1.

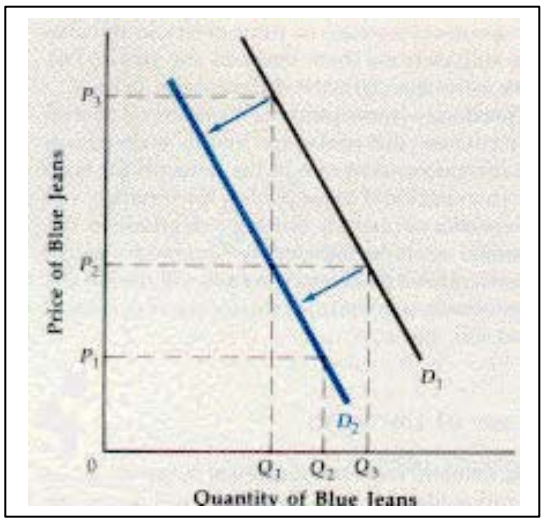


FIGURE 8.8 Decrease in Demand

A downward shift in demand, from D_1 to D_2, represents a decrease in the quantity of blue jeans consumers are willing to buy at each and every price. It also indicates a decrease in the price they are willing to pay for each and every quantity of jeans. At price P_2, for instance, consumers will now buy only Q_1 jeans (not Q_3, as before); and they will now pay only P_2 for Q_1 jeans -- not P_3, as before.

Thus, while economists can confidently predict the directional movement of consumption when prices change, they cannot say what will happen to the demand for a particular good when income changes, because each individual determines whether a particular good is a normal, inferior, or luxury good. Different people will tend to answer this question differently in different markets. Beans may be an inferior good to most low-income consumers and a normal good to many others.

For example, how do you think a change in income will affect the demand for low-, medium-, and high-quality liquor? You may have some intuitive notion about the effect, but you are probably not as confident about it as you are about the effect of a price

decrease. In fact, during past recessions, the demand for both low- and high-quality liquor has increased. Some consumers may have switched to high-quality liquor to impress their friends, and to suggest that they have been unaffected by the economic malaise. Others may have tried to maintain their old level of consumption by switching to a low-quality brand.

The effect of a change in the price of other goods is similarly complicated. Here the important factor is the relationship of one good—say, ice cream—to other commodities. Are the goods in question substitutes for ice cream, like frozen yogurt? Are they complements, like cones? Are they used independently of ice cream? Demand for ice cream is unlikely to be affected by a drop in the price of baby rattles, but it may well decline if the price of frozen yogurt drops.

Two products are generally considered substitutes if the demand for one goes up when the price of the other rises. The price of a product does not have to rise above the price of its substitute before the demand for the substitute is affected. Assume that the price of sirloin steak is $6 per pound and the price of hamburger is $2 per pound. The price difference reflects the fact that consumers believe the two meats are of different quality. If the price of hamburger rises to $4 per pound while the price of sirloin remains constant at $6, many buyers will increase their demand for steak. The perceived difference in quality now outweighs the difference in price.

Because complementary products—razors and razor blades, oil and oil filters, VCRs and videocassette tapes—are consumed jointly, a change in the price of one will cause an increase or decrease in the demand for both products at once. An increase in the price of razor blades, for instance, will induce some people to switch to electric razors, causing a decrease in the quantity of razor blades demanded and a decrease in the demand for safety razors. Again, economists cannot predict how many people will decide the switch is worthwhile, they can merely predict from theory the direction in which demand for the product will move.

Derivation of Demand from Indifference Curves And the Budget Line

Our discussion of theoretical foundations of demand has, admittedly, been casual. Here we can add greater precision to the analysis. Much of the discussion has been founded on the notion of the rational pursuit of individual preferences. That is, we assume the individual knows what he or she wants and will seek to accomplish those goals. Preference, however, is a nebulous concept. To lend concreteness to the idea, economists have developed the indifference curve.

Individuals face limits in what they can produce and buy, a point of earlier chapters. That fact, together with the existence of indifference curves, can be used to derive an individual's demand for a product.

Derivation of the Indifference Curve

Consider a student whose wants include only two goods, pens and books. Figure 8.9 shows all the possible combinations of pens and books she may choose. The student will prefer a combination far from the origin to one closer in. At point *b*, for instance, she will have more books and more pens than at point *a*. For the same reason, she will prefer *a* to *c*. In fact the student will prefer *a* to any point in the lower left quadrant of the graph and will prefer any point in the upper right quadrant to *a*.

We can also reason that the student would prefer *a* to *d*, where she gets the same number of pens but fewer books than at *a*. Likewise, she will prefer *e* to *a* because it yields the same number of books and more pens than *a*. If *a* is preferred to *d* and *e* is preferred to *a*, then as the student moves from *d* to *e*, she must move from a less preferable to a more preferable position with respect to *a*. At some point along that path, the student will reach a combination of books and pens that equals the value of point *a*. Assuming that combination is *f* (it can be any point between *d* and *e*), we can say that the individual is indifferent between *a* and *f*.

Using a similar line of logic, we can locate another point along the line *gih* that will be equal in value to *a* and therefore to *f*. In fact, any number of points in the lower right-hand and upper left-hand quadrants of the graph are of equal value to *a*. Taken together, these points form what is called an indifference curve (see curve I_1 in Figure 8.10).

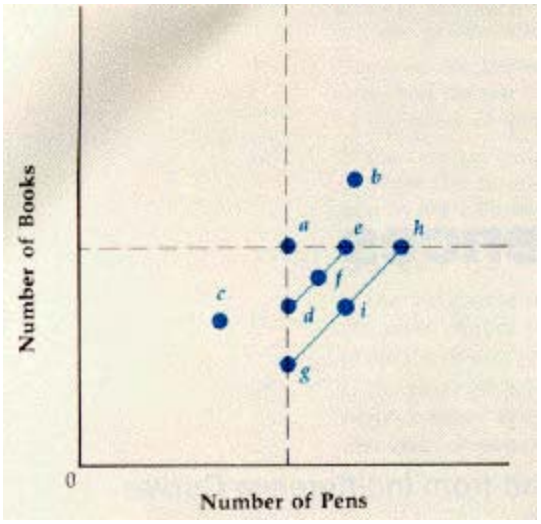


FIGURE 8.9 Derivation of an Indifference Curve

Because the consumer prefers more of a good to less, point *a* is preferable to point *c*, and point *b* is preferable to point *a*. If *a* is preferable to demand but *e* is preferable to *a*, then when we move from point *d* to *e*, we must move from a combination that is less preferred the one that is more preferred. In doing so we must cross a point—for example, *f*—that is equal in value to *a*. Indifference curves are composed by connecting all those points—*a*, *f*, *i*, and so on—that are of equal value to the consumer.

Using a similar line of logic, we can locate another point along the line *gih* that will be equal in value to *a* and therefore to *f*. In fact, any number of points in the lower right-hand and upper left-hand quadrants of the graph are of equal value to *a*. Taken together, these points form what is called an indifference curve (see curve I_1 in Figure 8.10). An **indifference curve** shows the various combinations of two goods that yield the same level of total utility.

Using the same line of reasoning, we can construct a second indifference curve through point *b*. Because *b* is preferable to *a*, and all points on the new indifference curve will be equal in value to point *b*, we can conclude that any point along the new curve I_2 is preferable to any point on I_1. Using this same procedure, we can continue to derive any number of curves, each one higher than, and preferable to, the last.

From this line of reasoning, an economist can draw several conclusions about the student's preference structure (called an "indifference map"):

1. The student's total utility level rises as she moves up and to the right, from one indifference curve to the next.
2. Indifference curves slope downward to the right.
3. Indifference curves cannot intersect. (An intersection would imply that all points on all the intersecting curves are of equal value, contradicting the conclusion that higher indifference curves represent higher levels of utility).

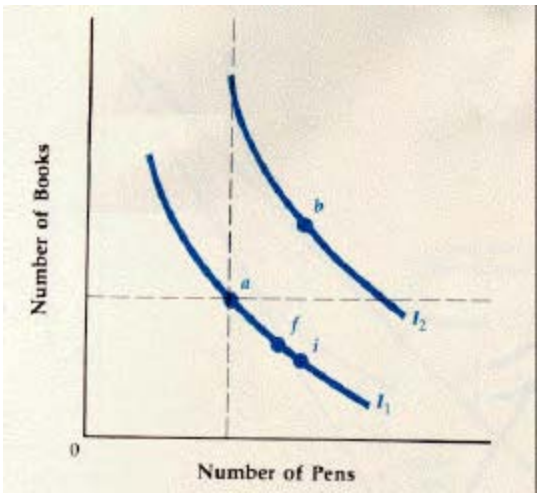


FIGURE 8.10 Indifference Curves for Pens and Books

Any combination of pens and books that falls along curve I_1 will yield the same level of utility as any other combination on that curve. The consumer is indifferent among them. By extension, any combination on curve I_2 will be preferable to any combination on curve I_1.

The Budget Line and Consumer Equilibrium

From indifference curves we can derive the law of demand. First we need to construct the individual's budget line, a special form of the production possibilities curve. The budget line shows graphically all the combinations of two goods that a consumer can buy with a given amount of income. Assume that our student earns an income of $150, which she uses to buy books and pens. Books cost $3 each and pens cost $5 a package. The student can spend all $150 on fifty books or thirty pen packs, or she can divide her expenditures in any number of ways to yield various combinations of books and pens. By plotting all the possible combinations, we obtain the student's budget line, B_1 P_1 in Figure 8.11.

All combinations on the budget line are possible for the student. She can choose point *a*, twenty-five books and fifteen pen packs, or point *b*, forty-five books and three

pen packs. Either combination exhausts her $150 budget. The rational individual will choose that point where the budget line just touches (is tangent to) an indifference curve—point *a* in this case.[4] Points farther up or down the budget line will put the student on a lower indifference curve and are therefore less preferable. (If, for instance, the student moves to *c* on the budget line, she will be on a lower indifference curve, I_2 instead of I_1.) At point *a*, the individual's wants are said to be in equilibrium. As long as her income and preferences and the prices of books and pens remain the same, she has no reason to move from that point.[5]

---

[4] This tangency condition can be derived mathematically by maximizing the consumer's utility subject to the budget constraint, or by maximizing the $U(X,Y)$ with respect to X and Y, subject to $P_xX + P_yY = I$. This constrained maximization problem can be carried out by forming the Lagrangian function

$$L = U(X_1Y) + l(I - P_X X - P_Y Y)$$

where λ is known as a *Lagrangian multiplier*, and maximizing it with respect to X and Y and minimizing it with respect to λ. The necessary conditions are

$$\frac{\partial L}{\partial X} = \frac{\partial U}{\partial X} - l P_X = 0$$

$$\frac{\partial L}{\partial Y} = \frac{\partial U}{\partial Y} - l P_Y = 0$$

$$\frac{\partial L}{\partial l} = I - P_X X - P_Y Y = 0$$

Equation (1) can be divided by equation (2), which, after simple algebraic manipulation, yields

(Missing equation to be added).

The left-hand side of this equation is -1 multiplied by the ratio of the marginal utility of good X to the marginal utility of good Y, or the slope of the indifference curve. The right-hand side is -1 multiplied by the ratio of the price of good X to the price of good Y, or the slope of the budget constraint.

The equality of these two slopes is dependent on the assumption that the consumer will consume positive quantities of both goods. Later in this chapter, we will consider the possibility that the consumer may maximize utility subject to the budget constraint by deciding to consume none of one of the goods.

[5] We can provide another intuitive rationale for the required condition for consumer equilibrium. Starting with the tangency requirement

$$\frac{MU_X}{MU_Y} = \frac{P_X}{P_Y}$$

we can obtain the equivalent condition

$$\frac{MU_X}{P_X} = \frac{MU_Y}{P_Y}$$

by simple algebraic manipulation. Verbally, this means that the consumer receives the same increase in utility from spending $1 more on good X as would be received from spending more on good Y. We can see that this condition is necessary if utility is being maximized subject to the budget constraint by assuming that the condition is not satisfied. Assume for example, that (continued on next page)

$$\frac{MU_X}{P_X} > \frac{MU_Y}{P_Y}$$

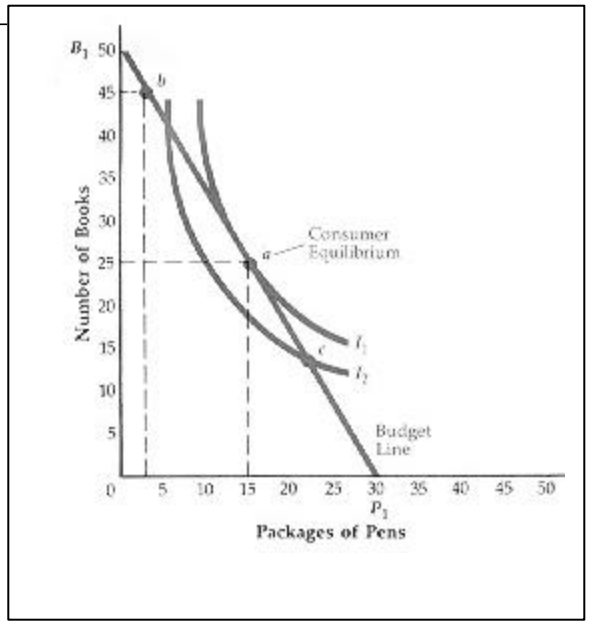


FIGURE 8.11 The Budget Line and Consumer Equilibrium

Constrained by her budget, the consumer will seek to maximize her utility by consuming at the point where her budget line is tangent to an indifference curve. Here the consumer chooses point *a*, where her budget line just touches indifference curve I_1. All other combinations on the consumer's budget line will fall on a lower indifference curve, providing less utility. Point *c*, for instance, falls on indifference curve I_2.

What happens if prices change? Suppose the individual's wants are in equilibrium at point *a* in Figure 8.11 when the price of pens falls from \$5 a pack to \$3 a pack. (The price of books stays the same.) The budget line will pivot to B_1P_2 in Figure 8.12, reflecting the greater buying power of the student's income. (She can now buy fifty pen packs with \$150.) The new budget line gives the student a chance to move to a higher indifference curve—for instance, to point *c*, twenty-two pens and twenty-eight books.

The Law of Demand, Again

The result of the price reduction is that the student buys more pens. Thus we derive the law of demand, that quantity demanded is inversely related to price. The downward-sloping demand curve for pens shown in Figure 8.13 is obtained by plotting the quantities of pen packs bought from Figure 8.12 against the price paid per pack. When the price of pens falls from \$5 to \$3 a pack in Figure 8.12, the consumer increases the quantity purchased from fifteen to twenty-two packages.

This tells us that if \$1 less is spent on good Y, utility will not decline as much as it will increase if \$1 more is spent on good X. Therefore, the consumer can increase total utility without increasing expenditures by reducing the consumption of good Y and increasing the consumption of good X. This will continue to be true until the equality is restored, which will happen eventually as MU_Y increases relative to MU_X. In a similar manner, we can argue that the consumer will move toward the equilibrium condition if we assume that

$$\frac{MU_X}{P_X} < \frac{MU_Y}{P_Y}$$

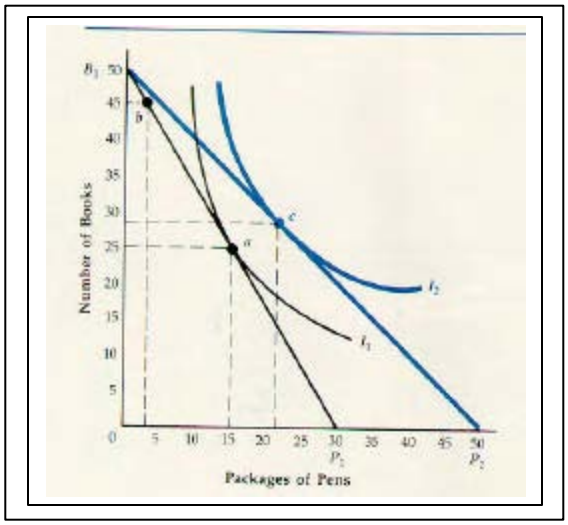


FIGURE 8.12 Effect of a change in Price on Consumer Equilibrium

If the price of pens falls, the consumer's budget line will pivot outward, from B_1P_1 to B_1P_2. As a result, the consumers can move to a higher indifference curve, I_2 instead of I_1. At the new price the consumer buys more pens, twenty-two packs as opposed to fifteen.

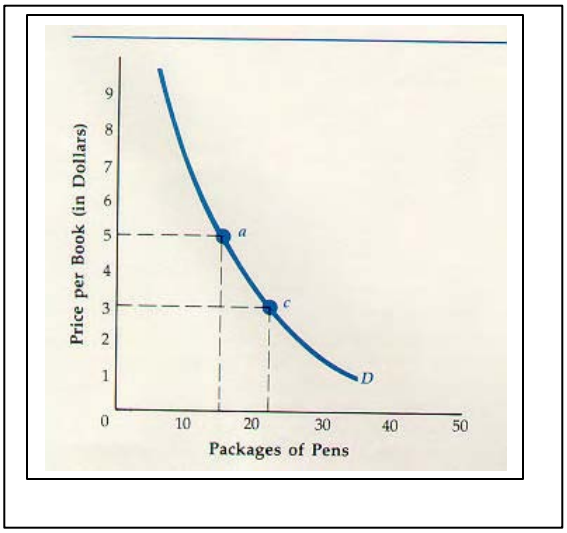


FIGURE 8.13 Derivation of the Demand Curve for Pens

When the price of pens changes, shifting the consumer's budget line from I_1 to I_2 in Figure 8.14, the consumer equilibrium point changes with it. The consumer's demand curve for pens is obtained by plotting her equilibrium quantity of pens at various prices. At \$5 a pack, the consumer buys fifteen packs of pens (point *a*). At \$3 a pack, she buys twenty-two packages (point *c*).

Application: Cash Versus In-Kind Transfers

A cash grant will raise the welfare of the poor more than an in-kind transfer of equal value. Figure 8.14 illustrates a poor family's budget line for higher education and housing, H_3 E_3. Without subsidies, this family can buy as much as E_3 units of education (and no housing) or H_3 units of housing (and no education). Because the family wants both housing and higher education, it will probably divide its income between the two, choosing some combination like point *a*, or E_1 education and H_1 housing.

Suppose that the government decides to subsidize the family's higher education purchases through reduced university tuition. Its action lowers the total price of education, pivoting the family's budget line out to H_3E_5. The result is that the family can now consume more of both items, education and housing. The family will probably move to some combination like *b*, H_2 housing and D_2 education Its education consumption has gone up and the additional housing purchased represents an increase in income equal to the vertical distance between *b* and *c*.

Suppose the family were given the cash equivalent of *bc* instead. The additional money would not change the relative prices of higher education and housing, as the reduced tuition program did. It would shift the budget line from H_3E_3 to a parallel position, H_4E_4 (dashed line). The relative price of housing is lower on H_4E_4 than on

H_3E_5. Thus the family would tend to prefer *d* to *b*, both of which are available on line H_4E_4, we must presume that they would prefer cash to an in-kind subsidy.

This point can be seen even more clearly with the help of indifference curves. Imagine an indifference curve tangent to H_3E_3 in the absence of government relief, causing the family to point *a*. Imagine a higher indifference curve that is tangent to H_3E_5 at point *b*. Now, imagine an even higher indifference curve tangent to H_4E_4 at point *d*.

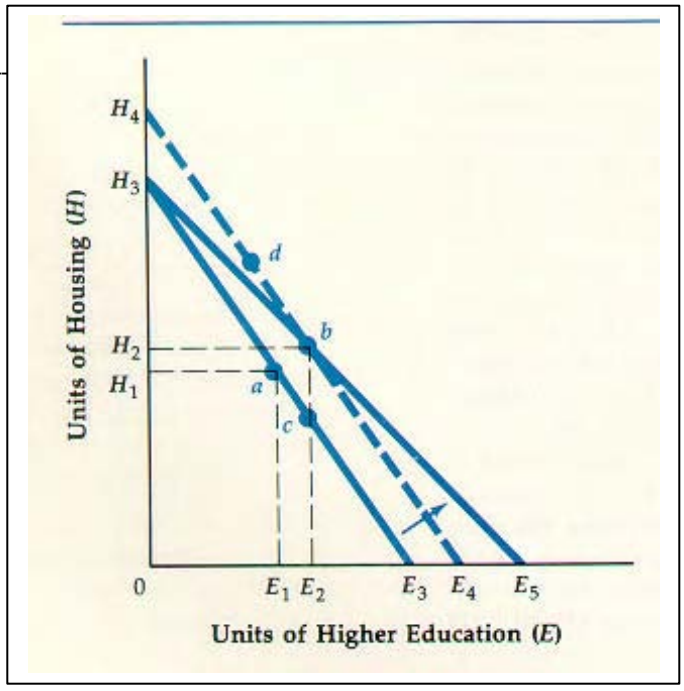


FIGURE 8.14 Budget Line: Cash Grants versus Food Stamps

If the price of education is reduced by an in-kind subsidy, a family's budget line will pivot from H_3E_3 to H_3E_5. The family will move from point *a* to point *b*, where it can consume more food and housing. If the family is given the same subsidy in cash, its budget line will move from H_3E_3 to H_4E_4. Since the relative price of housing is lower on H_4E_4 than on H_3E_5, the family will choose a point like *d* over *b*. Since *b* was the family's preferred point on H_3E_5, but they prefer *d* to *b*, we must presume they also prefer cash to a food subsidy.

Application: Capturing the Consumer surplus

The price that a consumer pays for a good reflects the value that he or she places on an additional unit of the good. Since the price normally applies uniformly to all units of the good purchased and the consumer generally values the last unit consumed less than the units consumed previously, the consumer values the total consumption of a good at more than the amount paid for its consumption. The gap between what a consumer is willing to pay rather than do without a good (the *total value* placed on the good) and what the consumer actually pays is referred to as the **consumer surplus**. Obviously, suppliers prefer that consumers pay more rather than less for a good and are anxious to capture as much consumer surplus as possible. We can employ indifference-curve analysis to show how suppliers use different pricing schemes to encourage consumers to pay more for a given quantity of a good than they would if the good were uniformly priced.

Conceptually, the simplest way for a supplier to capture the total consumer surplus of an individual would be to charge a different price for each unit consumed and to price each unit at the maximum amount the consumer is willing to pay for that unit. But such a pricing policy would be enormously difficult to implement. The supplier

would have to obtain detailed information about all consumers' preferences. Also, consumers who place a relatively small value on the good and therefore purchase it for less, would have to be prevented from selling the good to consumers who value it more highly. Otherwise, low-demand consumers would be able to buy the good at a relatively cheap price and profitably undercut the price that the supplier is charging the high-demand consumers.

A final and related difficulty is that the more competitors a supplier has, the more difficult it is to charge the same customer different prices for different units or to charge different customers different prices. Although a consumer may be willing to pay the only supplier of a good more for the first unit than the second unit, more for the second unit than the third unit, and so on, this is not necessarily true when the consumer can choose among several suppliers. A consumer will not be willing to pay one supplier any more for a particular unit of a good than is being charged by an alternative supplier. The consumer still values the first unit of the good more than the second unit, but competition among several suppliers makes it difficult for any one supplier to take advantage of this fact by imposing a different pricing strategy on each consumer and charging each consumer a different price for each unit purchased. However, relatively crude or simple price-discrimination schemes can be implemented that allow suppliers to capture more of the consumer surplus than they could under a uniform pricing policy.

Such a price-discrimination scheme is illustrated in Figure 8.15 with the aid of an individual's indifference-curve map. We assume the individual is initially at $\overline{Y}$, consuming $\overline{Y}$ units of good Y and no units of good X. Indifference curve I_1 indicates the consumer's level of satisfaction for this consumption bundle. Given an opportunity to purchase good X at a uniform price, reflected in the slope of budget constraint $\overline{YX}$, the consumer will purchase X_2 units and increase satisfaction by reaching the higher indifference curve I_2. Increased satisfaction is derived because less is being paid for the X_2 units than they are worth to the consumer. The total value of the X_2 units to the consumer is given in Figure 8.15 by the distance $\overline{Y} - Y_1$, which is the maximum amount of good Y the consumer is willing to sacrifice to obtain X_2 units. (The consumer is indifferent between $\overline{Y}$ units of Y and no X and Y_1 units of Y and X_2 units of X.) But given the budget constraint $\overline{YX}$, the consumer only has to sacrifice $\overline{Y} - Y_2$ units of Y to obtain X_2 units of X. The distance $Y_2 - Y_1$ measures the consumer surplus associated with the consumer's ability to buy good X at a uniform price.

The supplier is interested in whether a relatively simple pricing strategy will capture some of this consumer surplus. The supplier's objective is to raise the price and still have the consumer purchase the same quantity of good X. If the supplier raises the price uniformly, however, the budget constraint will pivot to the left around point $\overline{Y}$, and the consumer can be expected to purchase fewer units of good X. But what will happen if the supplier imposes a **two-part pricing policy**, which allows the consumer to purchase good X at a lower price if a specified number of initial units of X are purchased at a higher price? Assume, for example, that the consumer faces the budget constraint $\overline{Ya}\overline{\overline{X}}$ in Figure 8.15. If the first X_1 units of good X are purchased at the higher price

reflected by the budget constraint segment $\overline{Y}a$, then the consumer can buy additional units of good X at the much lower price indicated by the segment $a\overline{\overline{X}}$ of the budget constraint. Faced with such a pricing policy, the consumer will be willing to sacrifice $\overline{Y} - Y_1$ units of good Y, to buy X_2 units, thereby dissipating all of the consumer surplus.[6]

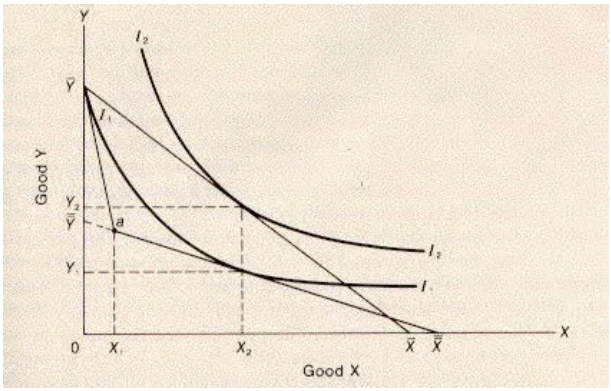


Figure 8.15. Two-Part Pricing

A uniform price can lead the individual to buy Y_2 and X_2. However, a two-part price can lead the individual to buy the same amount of X while reducing the purchases of Y to T_1, leaving the consumer on a lower utility curve and the seller with more income.

Two-part pricing strategies in the real world are not usually calibrated accurately enough to capture an individual's entire consumer surplus. Also, the same two-part pricing policy normally applies to everyone, even though preferences – and therefore indifference curves – vary from consumer to consumer. Thus, any given two-part pricing strategy will capture more consumer surplus from some than from others. However, such a strategy generally allows suppliers to motivate consumers to pay more for a given quantity of a good than they would under a uniform pricing policy.

Given the advantage that suppliers can realize from a two-part pricing strategy, it is not surprising that different variations of such pricing strategies are often encountered. For example, suppliers of electricity almost universally employ at least a two-part pricing schedule, so that the first few kilowatts of power used during the billing period cost the consumer more than subsequent kilowatts. A variation on two-part pricing is the membership fee – an initial charge that entitles the consumer to purchase a product at a lower price. As shown in Figure 8.15, this produces the same effect as straight two-part pricing. Assume that on paying an initial fee of $\overline{Y} - \overline{\overline{Y}}$, the consumer can buy all the units of X desired at the reduced price, reflected in the budget constraint $\overline{\overline{Y}}\overline{\overline{X}}$. We can see that the consumer will respond to this pricing policy by paying the fee and purchasing X_2 units of good X, allowing the supplier to capture all the consumer surplus. Automobile-rental firms use a form of this pricing policy when they impose a daily charge plus a per-mile charge. Computer time is commonly obtained by paying a lump-sum rental, which then entitles the individual to use the computer at a low hourly charge. Amusement parks

[6] Actually, the consumer is indifferent between buying no X and buying X_2 units of X. But if the consumer buys any of good X at all, it will be X_2 units, and only the slightest decrease in the price of good X along either segment of the budget constraint will make the purchase of X the most attractive alternative.

usually charge an entry fee and then attach no marginal charge to the rides. Surely you can think of other examples of two-part pricing policies.

Application: Charity Versus Corner Solutions

The underlying assumption that individuals are motivated to maximize their own utility may leave the impression that there is no room in our analysis for concern for others. This is not true. The indifference-curve approach to utility maximization can be used to explain *charitable behavior*.

Nothing in our analysis prevents an individual's utility from being influenced by the consumption of *others* as well as by his or her own consumption. For example, let's assume that we are considering two individuals – individual D (the donor) and individual R (the recipient) – and that D's utility is a function not only of his own consumption but of R's as well. D's preferences can be expressed with indifference curves showing combinations of D's and R's consumption that provides D with the same utility. Two such indifference curves are shown in Figure 8.16. These curves indicate that when D has a high income relative to R's income, D is willing to transfer come income to R. As expected, however, as D's income declines relative to R's income, the slope of the indifference curve becomes shallower, indicating that D is willing to sacrifice less income to increase R's income by an additional dollar. And if R's income increases too much relative to D's income, envy sets in and individual R's income becomes a "bad" (a "good" with negative value) to D. This is shown by the upward-sloping portions of the indifference curves. Once envy sets in, D's income will have to be increased before he is willing for R's income to increase.

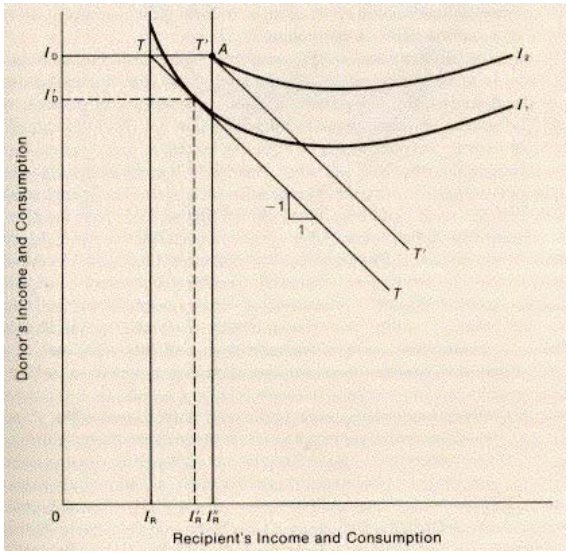


Figure 8.16 Sometimes It Is Better to Give Than to Receive

The donor, D, has an initial starting income that is higher than the recipient's, R's. By giving income TT' to R, D moves to a higher indifference curve. This is a case in which R's welfare affects D's, leaving D better off by giving than receiving.

Now let's assume that D can transfer income costlessly to R, or that R receives an additional dollar for each dollar D gives up. This is reflected in Figure 8.16 by – 1 slope of *TT*, which shows the different income combinations that D can realize by transferring

income to R. (Here I_D and I_R represent D's and R's initial incomes, respectively.) Subject to this constraint, D attempts to maximize his utility through charitable contributions, reaching indifference curve I_1 by donating $I_D - I_D'$ dollars to R. This increases R's income from I_R to I_R'.

Next we will assume that R's income increases to I_R'' without any transfers from D. The relevant constraint D faces in donating income to R is now given by *T'T'* in Figure 8.16. But with this constraint, D maximizes utility by *not* donating any income to R. At point A, the constraint is steeper than the indifference curve, resulting in a corner solution. D does not donate any income increase to R: the first dollar that D donated would increase R's income by a lesser amount than is required to make D willing to sacrifice $1 of income.

Application: Charity and Paternalism

Due to an underlying fear that the recipients will not spend the money in their best interests, few organized charities, either public or private, simply transfer income to the needy. Instead of money, **charitable contributions** normally consist of particular goods and services that the donors believe the recipients should have. It will be helpful to use the indifference-curve approach to consumer behavior to analyze the effect of these **in-kind gifts** in terms of the intent of the donors and the utility of the recipients.

The three indifference curves I_1, I_2, and I_3 in Figure 8.17 belong to an individual who is to be the recipient of a donated good – say, bus transportation. Before the donation, the individual's budget constraint with respect to bus transportation and all other goods is defined by line *BC*. Given this constraint, the individual will maximize utility by choosing bundle A (point A) and consuming bus rides at the rate of X_1 per week. Now we will assume that this individual qualifies for public relief, which takes the form of free bus transportation – something the transit authorities feel people should be encouraged to consume. Letting $\overline{X}$ be the quantity of free bus transportation received, the budget constraint becomes *BDE*. Beyond point *D*, the slope of this budget constraint is the same as *BC*, reflecting the fact that the regular price must be paid for bus transportation in excess of $\overline{X}$. Faced with this new budget constraint, the consumer will maximize utility by choosing bundle D, point *D* at the kink in the constraint. Consumption of bus transportation will increase from X_1 to $\overline{X}$, and utility will also increase because the individual moves from indifference curve I_1 to I_2.

Two objectives have been accomplished by this contribution. First, the recipient's well being was improved; second, the recipient's consumption of bus transit was increased – something those controlling the contribution thought was important. It is worth noting that these two objectives are somewhat in conflict with one another. For example, if the only objective had been to increase the recipient's well being as much as possible, the contribution would have been made in the form of money or some other form of general purchasing power. If instead of $\overline{X}$ bus tokens, the recipient had received enough *money* to buy $\overline{X}$ bus tokens, then the budget constraint would have been *FDE*. Given this constraint, the individual would have chosen bundle G at point *G*, consumed

only X_2 units of bus transportation, and reached indifference curve I_3, thereby attaining a higher utility level than that achieved when the in-kind gift of bus rides was given. Of course, this increase in utility would have been attained at the expense of the people who felt that the individual actually needed $\overline{X}$ units of bus transportation.

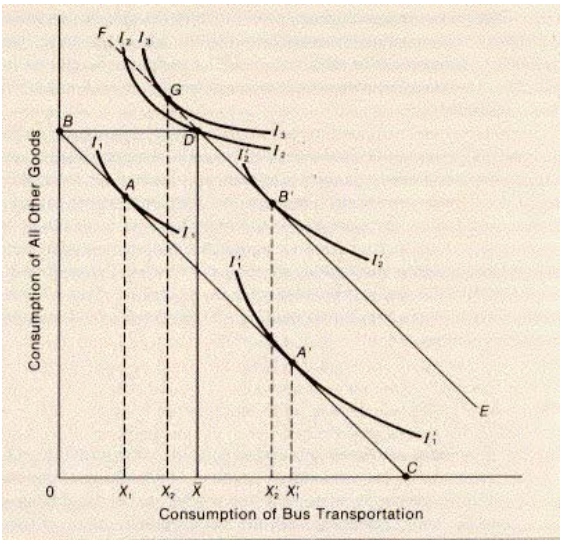


Figure 8.17 In-Kind Charitable Contributions

An in-kind charitable contribution can lead to a person buying more of the good, as is the case when the individual moves from A to D (although he would prefer to move to G, but can't). However. The charitable contribution can also lead to the individual consuming less of the charitable good, which is what happens when the individual moves from A' to B'.

Next, we will consider an individual who, before the receipt of $\overline{X}$ units of free bus transportation, was consuming a greater quantity than that. The preferences of this individual are represented in Figure 8.17 by indifference curves I_1' and I_2'. Before the gift, *BC* is the budget constraint and the individual maximizes utility by choosing bundle A' at point *A*' and consuming X_1' units of bus transportation. But after the receipt of $\overline{X}$ units of bus transit, the budget constraint shifts to *BDE* and the consumer maximizes utility by choosing bundle B' and point *B*' and reducing consumption to X_2' units of bus transportation. The individual reduced bus-service consumption when this service was given free of charge. This can only occur if the individual regards bus service as an inferior good, which most people do. The gift increases the recipient's real income and motivates a reduction in the consumption of an inferior good as long as the relative price of that good remains constant. And since we assume that the individual was consuming more than $\overline{X}$ before the gift, the relative price of the marginal unit consumed is not affected by the gift. The effect of giving the recipient $\overline{X}$ units of bus transportation is the same as giving the individual enough money to purchase that much bus service.

This analysis can be applied to the current food-stamp program in the United States. As that program is now structured, people who qualify are given food stamps in specific dollar amounts that can be redeemed for food. If the dollar value of the stamps exceeds the amount the recipients are spending on food, then the program can be expected to motivate a larger increase in food consumption than would result from an equivalent income transfer. But if more money was being spent on food before the program was initiated than the dollar value of the food stamps received after the program

was begun, then there is no effective difference between providing recipients with food stamps or with equivalent amounts of cash.

The Demand for Public Goods

Early in the book we distinguished two types of goods, private and public goods. To this point we have developed the market demand for a private good. The development of a community's demand for a *public (or community) good* is substantially different from the previous construction of the demand for a private good. As we will see, the nature of a public good prevents its provision by private firms in an efficient manner.

We can construct the demand for a public good by first noting that each individual has a downward sloping demand for public goods -- national defense, environmental quality, etc. However, if a unit of a public good is provided, all within the relevant group can receive benefits from it. This is not true of a private good; a unit of a private good benefits only the person who possesses it. Consequently, the *community's* demand for a public good must be obtained by vertically adding up the values (as measured by the price) that all within the group place on each unit. The reason for this is simply that all can benefit from each unit; and the relevant question is what is the total value, which all within the group place on all units. This is why we vertically add each unit, to find the total value.

To illustrate in the simplest possible terms, consider a community made up of only two people. There is some good, like police protection, from which both can receive benefits simultaneously. Suppose, however, that individual B has a greater demand for police protection than A. This condition is illustrated in Figure 8.18. For the first unit of police protection, A is willing to pay as much as \$3, which is an indication of the relative value he places on that unit. B, on the other hand, is willing to pay more, \$5 in this example. Both can benefit from the first unit of police protection, and the collective value attached to this unit is \$8 (\$3 + \$5). Since each individual's demand curve is downward sloping, the relative value attached to each unit declines as the quantity is increased. For the second unit, A is willing to pay \$2 and B, \$4. Collectively, they are willing to pay as much as \$6. For the fourth unit, A is unwilling to pay anything; however, B is still willing to pay as much as \$2. From that point on, the collective value of police protection is simply equal to the value which B places on the good.

The prices which A and B are willing to pay for each unit are shown and added together in Table 8.2. If we plot the collective value which A and B place on each unit (P_{A+B} in the table) against the quantity, we will obtain a demand curve represented by the darker line in Figure 8.18. This curve represents the demand for a public good or service. It represents what the people in the community are willing to pay in total (if they have to) for each unit of police protection. We talk in terms of the collective value of the community because all people can share in the benefits of the good or service if it is provided.

Public officials may be unable to obtain a very accurate picture of the public's demand for a service like police protection. When people vote for representatives or in a referendum, they vote for or against or yes or no. Their votes are only a crude indication

of the *intensity* of their preferences for the public good. This is so because all votes count the same. The politicians are forced to try to sense the mood of the people, and in doing that they may provide many people with an opportunity to misrepresent their true demand for the good, that is, how much they are willing to pay. More will be said on this in later chapters. The important point to understand at this juncture is the difference in the construction of the demand for public and private goods.

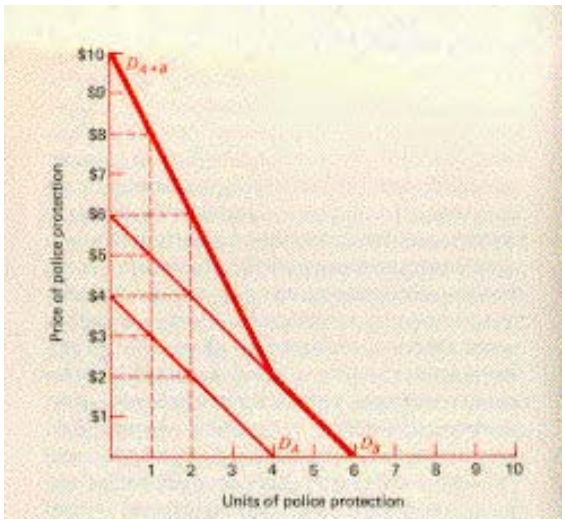


FIGURE 8.18 The Public Good Demand Curve

The public good demand curve, D_{A+B}, is equal to the vertical summation of the demands of the individuals A and B. The curves are added vertically because both individuals A and B can simultaneously benefit from each unit of police protection provided.

TABLE 8.2 Construction of a Public Goods Demand Curve

Units of Police Protection (1)	**Price A Is Willing To Pay for Each Unit (DA) (2)**	**Price B Is Willing To Pay for Each Unit (DB) (3)**	**Public Goods Demand: Price A and B Are Willing to Pay for Each Unit When They Act Collectively (DA+B) (4)**
0.5	$3.50	$5.50	$9
1	3	5	8
2	2	4	6
3	1	3	4
4	0	2	2
5	0	1	1
6	0	0	0

Note: The Public Goods demand curve in Figure 8.9 (the darker line) is obtained by plotting the prices that A and B are collectively willing to pay for each unit in column (4) against their respective units in column (1).

Irrationality and the Law of Demand

So far we have been discussing demand in terms of rational behavior. Even if some consumers behave irrationally, the law of demand will apply. As long as some people in the market respond rationally, demand will change with a change in price. For instance, many people buy cigarettes because they are addicted to them. At times habitual smokers may not consider price in making their purchases. Therefore we cannot expect the quantity they buy always to vary with price (except to the extent that it affects their total purchasing power). If occasional smokers take price into consideration when they buy, however, their demand for cigarettes will produce the normal downward sloping curve. If we add the quantity bought by smokers who are addicted to the quantity bought by those who are not, the total market demand curve will slope downward (see Figure 8.19). At a price of P_1, Q_1 cigarettes will be bought by addicted consumers, and $Q_3 - Q_1$ cigarettes will be bought by occasional consumers. If the price then rises to P_2, the total quantity bought will fall to Q_2, reflecting a predictable drop in the quantity purchased by occasional consumers.

This kind of reasoning can be extended to impulse buying. Some people respond more to the packaging and display of products than to their price. Their demand may not slope downward. As long as some people check prices and resist advertising, however, the total demand for any good will slope downward. Store managers must therefore assume that changes in price will affect the quantity demanded. The fact that some people may behave irrationally reduces the elasticity of demand but does not invalidate the concept of demand.[7]

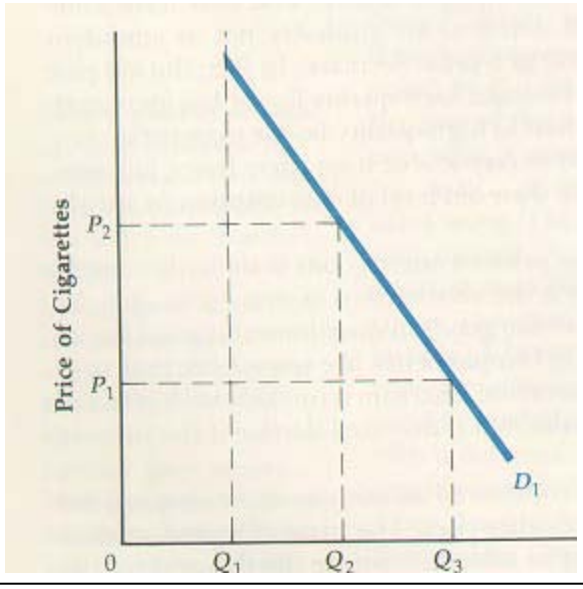


FIGURE 8.19 Demand Including Irrational Behavior

If irrational consumers demand Q_1 cigarettes no matter what the price, but rational consumers take price into consideration, market demand will be D_1. The quantity purchased will still vary inversely with the price.

[7] In fact, one economist has demonstrated rather convincingly that the assumption of rational behavior is unnecessary to the construction of a downward-sloping demand curve. The curve, he argues, can emerge from completely random behavior on the part of consumers. Gary S. Becker, *Economic Theory* (New York: Alfred A Knopf, 1972), pp. 19—23.

Lagged Demands and Network Externalities

Almost all microeconomics textbook do what we have done with demand, they provide a lengthy discussion of the demand for a "standard" good. They will explain that the quantity of the good purchased will be related to the price of the good in question and a number of other considerations (such as weather, income, and the prices of other goods), as we have stressed. The lower the price of a candy bar, for example, the greater the quantity purchased, and vice versa. This inverse relationship between price and quantity is so revered in economics that it has a special label, the "law of demand." Nothing will be said by most textbooks about how the benefits received by any one candy bar buyer in one time period will affect the benefits received in subsequent time periods, or, rather, how the consumption level today will affect the demand in the future. Little or nothing will also be written about how the benefits (and demand) depend upon how many other people have bought candy bars, all of which is understandable. The benefit that one person gets from eating a candy bar in one time period does not materially affect the benefits received from eating another bar later and is also not materially affected by how many other people buy bars in the various time periods. People just buy and consume candy bars independent of one another, and couldn't care less about how much other people enjoy their candy bars.

This is not true for two special classes of goods called **lagged demand goods** and **network goods.** A lagged demand good is one in which consumption today affects consumption tomorrow (or future time periods). A lagged demand good has one defining feature: the greater the quantity purchased today, the greater the demand tomorrow. Good examples of lagged demand goods include cigarettes, alcohol, and street drugs, given that they tend to be addictive in consumption. As we will see, the theory of lagged demand is similar to the theory of "rational addiction," or the view that before consumption begins, people can rationally weigh the long-term costs and benefits, or pros and cons, of consuming goods that can be physically compelling in consumption.

A network good is a product or service the value of which to consumers depends intrinsically on how many other people buy the good. A network good has one defining feature: The greater the number of buyers, the greater the benefits most, if not all, buyers receive. These goods are said to exhibit "network effect" or "network externalities," which has been appropriately described by one economist as "a phenomenon in which the attractiveness of a product to customers increases with the use of that product by others."[8] Good examples of network goods include telephones, fax machines, and computer software. One person's telephone is useless someone else owns a phone, and the more people who buy phones, the greater the value of the phone is to everyone, because more people can be called.

[8] Franklin M. Fisher, Direct Testimony, U.S vs. Microsoft Corporation, Civil Action No. 98-1233 (TPJ), filed October 14, 1998, p. 15. (as downloaded from http://www.usdoj.gov/atr/cases/f2000/2057.pdf).

As you can see, lagged demand and network goods have much in common, the interconnectedness of consumption, which has important implications for pricing strategy.

Lagged Demands

One of the authors of this book, Lee, was involved in the development of the theory of lagged demands.[9] He and David Kreutzer have argued that, for some purposes, the demands that apply to a given product but that are evident in different time periods should be viewed as interdependent, with consumption in the future critically tied to consumption in the current time period, whatever the elasticity of the current demand (or whatever the technical capacity of consumers to respond to price changes in the current time period) might be. From this perspective, a lagged demand good is one in which the future good is a complement to the current good; they go together. According to Lee and Kreutzer,

> The crucial assumption behind our analysis is that lags exist in the demand for the resource; future demands are influenced by current availability. The demand for petroleum is clearly an example of such a lagged demand structure, with future demand for petroleum significantly influenced by investment decisions made in response to current availability.[10]

Hence, it follows that like all complements, the future demand for a product depends upon the current price for the same good. Behind such an obvious point lie important insights that might otherwise go unrecognized from the usual view of demand (and, as we will see, excise taxes and other policy topics).

As a consequence of the complementarity in consumption over time, firms faced with lagged demand have an incentive to lower their current price in order to stimulate future sales. They might even might charge a price in (or marginally lower their price toward) the inelastic range of their current demand curves, in spite of the fact that they lose current revenues from doing so, just so they can stimulate a greater future demand, which will permit them to raise their future prices and which can lead to greater generate profits in the future. This is true, of course, so long as the producers' rights to exploit future profits are not threatened.

What is interesting about this perspective is that under conditions of lagged demand, a cartel may form not with the intent of raising the group's current price, but with the intent of lowering the current price and expanding demand, and profits, in the future.[11]

---

[9]Dwight Lee and David Kreutzer, "Lagged Demand and a 'Perverse' Response to Threatened Property Rights," Economic Inquiry, vol. 20 (October 1982), pp. 579-588.

[10]Lee and Kreutzer, "Lagged Demand and Property Rights," p. 580.

[11]Such a cartel may also dissolve because of rampant cheating involving price increases, with all firms seeking to benefit from the greater demand stimulated by lower prices charged by other cartel members.

Also, it needs to be noted that the conventional treatment of demand, under which the demand tomorrow is unrelated to the consumption level today, holds that the potential for future threats to the stability of property rights could lead to "over-production" during the current time period. This is the case because if a firm – for example, an oil company – fears it will lose its property rights to its reserves, then it has an incentive to increase production and expand sales today. Never mind that the added supply oil might depress the current price. The oil firm can reason that if it doesn't pump the oil out of the ground in the short term, it will not have rights to the oil in the future.

For goods subject to the lagged demand phenomenon, any looming threat to property rights can cause some firms to do the opposite, reduce production of oil (or the exploitation of any other resource), hike the current price, and extract whatever profits remain. When its property rights are threatened, the firm no longer has an incentive to artificially suppress its current price in order to cultivate future demand.

Rational Addiction

Two economists from the University of Chicago, Gary Becker and Kevin Murphy, have developed a similar line of argument. The major difference is that their purpose was primarily to develop an economic theory of "addiction," which is a general concept also intended to suggest a tie between current and future consumption of a good or activity.[12] The tie-in, however, is physical (or maybe chemical) as in the case of cigarettes. People's future demand for smokes can be tied to their current consumption simply because of the body's chemical dependency on the intake of nicotine. As in the case of lagged demand goods, producers of addictive goods have an incentive to suppress the current price of their good – cigarettes – in order to stimulate the future demand for it. The lower the current price, the greater the future demand and the greater the future consumption.

This complementarity in consumption for an addictive (and lagged demand) good is illustrated in Figure 8.20. At price P_1 in the current time period, the consumption will be Q_1 in the current time period. However, because of that current consumption level, the demand in the future rises to D_2. At a price of P_2, current consumption rises to Q_2, but the future demand rises to D_3. You can imagine that at even lower prices, P_3, there will be some even higher demand curve, D_3, in the future time period. You can see in the illustration why firms have an incentive to lower the current price: the future demand rises. With other complement goods, if the price of one complement goes down and more of it is sold, then the demand for the other complement will go up, with its price rising. The same thing happens in this case. The only difference is that the complements are the same good but consumed in different time periods.

The current demand for one addictive good, cigarettes, might be highly inelastic, as is commonly presumed in microeconomics, but this does not mean that the long-run demand is necessarily inelastic. As illustrated in Figure 8.20, the short-term demand

[12]Becker and Murphy, "Rational Addiction."

curves (the dark lines) are each very inelastic, but the long-term demand curve (dashed line) is rather elastic. Indeed, Becker and Murphy maintain that the more addictive the good, the more elastic will be the long-term demand.[13] This is the case because a reduction in the *current* time period might not stimulate *current* sales very much. However, for highly addictive goods, current consumption can give an even greater increase in the future demand because the buyers "have to have more of it bad," thus resulting in even more future consumption than would be the case for less addictive goods. Hence, it is altogether understandable why cigarette firms decades ago would often have "cigarette girls" parading around campus in short skirts giving away small packs of cigarettes and why many drug dealers to this day eagerly give away the first "hits" to their potential customers. Indeed, it seems reasonable to conclude from the Becker/Murphy line of argument, the more addictive the good, the lower the current price. We might not even be surprised that for some highly addictive goods, the producers would "sell" their goods at below zero prices (or would pay their customers to take the good).

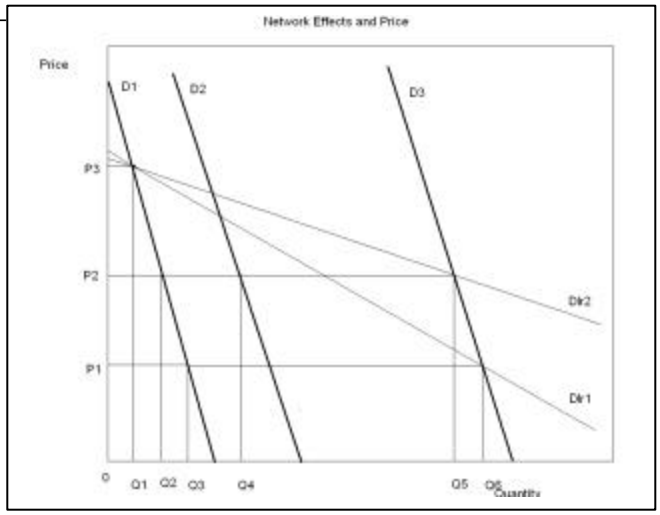


FIGURE 8.20 The Lagged Demand Curve

As the price falls from P_3 to P_2, the quantity demanded in the short run rises from Q_1 to Q_2. However, sales build on the sales, causing the demand in the future to expand outward to, say, D_2. The lower the price in the current time period, the greater the expansion of demand in the future. The more the demand expands over time in response to greater sales in the current time period, the more elastic in the long-run demand.

In contrast to the theory of lagged demand, this theory of rational addiction suggests explanations for a variety of behaviors, most notably, the observed differences in the consumption behavior of young and old, the tendency of overweight people to go on "crash diets" even when they may only want to lose a modest amount of weight, or alcoholics who become "teetotalers" when they decide to curtail their drinking. Old people may be less concerned about addictive behavior, everything else held constant, than the young. Old people simply have less to lose over time from addictions than younger people (given their shorter life expectancies). People who are addicted to food

[13] Becker and Murphy conclude, "Permanent changes in prices of addictive goods may have a modest short-run effect on the consumption of addictive goods. This could be the source of a general perception that addicts do not respond much to changes in price. However, we show that the long-run demand for addictive goods tends to be more elastic than the demand for nonaddictive goods" (Ibid, p. 695).

may rationally choose to drastically reduce their intake of food even though they may need to lose only a few pounds because their intake of food compels them to "over-consume." Similarly, alcoholics may "get on the wagon" in order to temper their future demands for booze because even a modest consumption level can have a snowballing effect, with a little consumption leading to more drinks, which can lead to even more.

Standard excise-tax theory suggests that producers' opposition to excise taxes should be tempered by the fact that the tax can be extensively passed onto the consumers in the form of a price increase (that must always be less than the tax itself). The theory of lagged demand suggests otherwise: producers of such goods have a substantial incentive to oppose the tax because of the elastic nature of their long-run demands. While they may be able to pass along a major share of the tax in the short run, they will not be able to do so in the long run.

Lagged Demands, Rational Addiction, And Excise Taxes

As we showed early in this book, an excise tax imposed on the production of a good can be expected to have several effects.

- First, the supply of the good will be curbed.
- Second, the price consumers pay will rise with the curtailment in supply.
- Third, the price received by producers after the tax will fall.

The difference between the price paid by consumers and price received by the producers equals the excise tax. (See Chapter 5 for a graphical presentation of these points.)

As you might imagine, the consequences of an excise tax for a good subject to a lagged demand or rational addiction are not exactly the same. The excise tax might indeed decrease the supply curve, as is the case of the standard good covered in Chapter 5. However, the impact on price and quantity sold will not likely be the same. This is because of the incentive the producers have to suppress the current price to stimulate future demand. When the prospects of the excise tax being enacted are evident to producers, they can be expected to raise their prices currently (before the tax is enacted). This means that the prospects of an excise tax can lead to a higher current price being received by producers, as well as a lower quantity sold (even without the excise tax in effect). When the tax is imposed, the reduction in quantity sold can be from two forces. First, the price increase caused by the excise tax. Second, the price increase caused by the prospects of the tax and the fact that the tax might be raised in the future.

Network Externalities

The theory of "network effects" or "network externalities" shares one key construct with the theory of lagged demand and rational addiction: the interconnectedness of demands. The interconnectedness in the theory of lagged demand and rational addiction is through time. The interconnectedness in the theory of network effects and externalities is across people and markets. The theory of network effects and

externalities is best understood in terms of telephone systems that actually form "networks," that is, are tied together with telephone lines (as well as microwave disks and satellites). No one would want to own a phone or buy telephone service if he or she were the only phone owner. There would be no one to call. However, if two people – A and B -- buy phones then each person has someone to call, and there are two pair-wise calls that can be made: A can call B, and B can also call A. As more and more people buy phones, the benefits of phone ownership escalate geometrically, given that there are progressively more people to call and even more possible pair-wise calls. If there are three phone owners – A, B, and C – then calls can be made in six pair-wise ways: A can call B or C, B can call A or C, and C can call A or B. If there are four phone owners, then there are 12 potential pair-wise calls; five phone owners, 20 potential pair-wise calls; 20 phone owners, 380, and so forth. If the network allows for conference calls, the count of the ways calls can be made quickly goes through the roof with the rise in the number of phone owners. It's important to remember that the benefits buyers garner from others joining the network can rise just from the *potential* to call others; they need not ever call all of the additional joiners. Neither of the authors ever expects to call every business in the country, but each author still gains from having the *opportunity* to call any of the businesses that have phones.

Accordingly, the demand for phones can be expected to rise with phone ownership. That is to say, the benefits from ownership go up as more people join the network. Hence, people should be willing to pay more for phones as the count of phone owners goes up. Some of the benefits of phone ownership are said to be "external" to the buyers of phones because people other than those who buy phones gain by the purchases (as was true in our study of public goods and external benefits studied in the last chapter). In more concrete terms, when one of the authors, Lee, buys a phone, then the other author, McKenzie, gains from Lee's purchase -- and McKenzie pays nothing for Lee's phone. For that matter, everyone who has a phone gains more opportunities to call as other people buy phones, or as the network expands (at least up to some point). The gains that others receive from Lee's or anyone else's purchase are "external" to Lee, hence are dubbed "external benefits" or, more to the point of this discussion, "network externalities."

In passing, we note that networks and network goods tend to turn one basic economic proposition on its head. There is a canon in economic theory that we have stressed from the start: As any good becomes scarcer, it becomes more valuable. In the case of network goods, just the opposite is true: as the good become more abundant, its value goes up.[14]

There are two basic problems that a phone company faces in building its network. First, the company has the initial problem of getting people to buy phones, given that at the start the benefits will be low. Second, if some of the benefits of buying a phone are "external" to the buyer, then each buyer's willingness to buy a phone can be impaired. How does the phone company build the network? One obvious solution is for the phone company to do what the producers in the theory of lagged demand and rational addiction

[14] See Kevin Kelly, New Rules for the New Economy (New York: Viking/Penguin Group, 1998), chap. 3.

do: "under price" (or subsidize) their products – phones -- or, at the extreme, give them away (or even pay people to install phones in their houses and offices).

Software Networks

The network effects in the software industry – for example, operating systems -- are similar but, of course, differ in detail from the network effects in the telephone industry. Indeed, the software developer may face more difficult problems, given that the software development must somehow get the computer users on one side of the market and application developers on the other side to join the network more or less together.

Few people, other than "geeks," are likely to buy an operating system without applications (for example, word processing programs or games) being available. If a producer of an operating system is only able to get a few consumers to buy and use its product, the demand for the operating system can be highly restricted. This will be the case because few firms producing applications will write for an operating system with a very limited number of users, given the prospects of few sales for their applications. However, the applications written for the operating system can be expected to grow with the number of people using the system. Why? Because the potential sales for applications will grow with the expansion in the installed base of computers using the operating system. If more applications are written for the operating system, then more people will want to buy and use the operating system – which can lead to a snowball effect: more sales, more applications, and even more sales in an ever expanding array of people connected to the operating system by way of the invisible "network."

As in the case of telephones, some of the benefits of purchases of the operating system (and applications) are "external" to the people who buy them. People who join the operating system network increase the benefits of all previous joiners, given that they have more people with whom they can share computers or share data and manuscripts. All joiners have the additional benefit of knowing that a greater number of operating system users can increase the likelihood of more applications from which they can choose. However, as in phone purchases, when the benefits are "external," potential users have an impaired demand for buying into the network. The greater the "external benefits," the greater buying resistance (or willingness to cover the operating system cost).

The network may grow slowly at the start, because people (both computer users and programmers) might be initially skeptical that any given operating system will be able to become a sizable network (and provide the "external benefits" that a large network can provide). However, as in the case of phones, "abundance" (not scarcity) can imply greater value for the software/operating system network.

As the network for a given operating system grows, more and more people will begin to believe that the operating system will become sizable, if not "dominant," which means that the network can grow at an escalating pace. As the network grows, there can be some "tipping point," beyond which the growth in the market for the operating system will take on a life of its own, that is, grow at an ever faster pace *because* it has grown at an ever faster pace. People will buy the operating system because everyone else is using

it (which can mean, it needs to be stressed, that the self-accelerating growth in buyers of one operating system can translate into the contraction of the market share for other operating systems). After the "tipping point" has been reached, the firm's eventual market dominance – and monopoly power -- is practically assured, according to the Justice Department.

This discussion might have relevance to the history of the dominance of the Apple and Microsoft operating systems. Before the introduction of the IBM personal computer, Apple was the dominant personal computer, running the CP/M operating system.[15] However, IBM and Microsoft developed their respective operating systems, PC-DOS and MS-DOS, in 1981. At that time, ninety percent of programs ran under some version of CP/M.[16] CP/M's market dominance was likely undermined by two important factors: First, CP/M was selling at the time for $240 a copy; DOS was introduced at $40.[17] Second, the dominance of IBM in the mainframe computer market could have indicated to many buyers that some version of DOS would eventually be the dominant operating system. In addition, Apple refused to "unbundle" its computer system: it insisted on selling its own operating system with the Macintosh (and later generation models), and at a price inflated by the restricted availability of Apple machines and operating systems.

Microsoft took a radically different approach: It got IBM to agree to allow it to license MS-DOS to other manufacturers and then did just that to all comers, presumably in the expectation that the competition among computer manufacturers on price and other attributes of personal computers would spread the use of computers – and, not incidentally, Microsoft's operating system. The expected "abundance" of MS-DOS systems led to an even greater demand for such systems, and to a lower demand for Apple systems. Many people started joining the Microsoft network, presumably, not always because they thought MS-DOS or Windows was a superior operating system to Apple's, but because any inferiority in the technical capabilities (if that were the case) would be offset by the benefits of the greater size network. Supposedly, as the network story might be told, there was a "tipping point" for Microsoft sometime in the late 1980s or early 1990s (possibly with the release of Windows 3.1) that caused Windows to take off, sending Apple into a market-share tailspin.

In 1998, the Justice Department took Microsoft to court for violation of the nation's antitrust laws. Among other charges, the Justice Department maintained that Microsoft was a monopolist, as evidenced by its dominant (90+ percent) market share in the operating system market, and that Microsoft was engaging in "predatory" pricing of its browser Internet Explorer. Microsoft had been giving away Internet Explorer with Windows 95 and had integrated Internet Explorer into Windows 98. The Justice Department claimed that the only reason Microsoft could possibly have had to offer Internet Explorer is to eliminate Netscape Navigator from the market. We can't settle

---

[15] David S. Evans, Albert Nichols, and Bernard Reddy, "The Rise and Fall of Leaders in Personal Computer Software," (Cambridge, Mass.: National Economic Research Associates, January 7, 1999), p. 4.

[16] Ibid.

[17] Ibid.

these issues here. All we can actually do is point out that the Justice Department starts its case against Microsoft with the claim that software markets are full of "network effects." While it might be true that Microsoft may have been engaging in predatory pricing, all we can say here is that it may also be true that Microsoft was responding to the dictates of "network effects," underpricing its product in order to build its network and future demand. It had another reason to lower its price to levels that Netscape might not consider reasonable. If Microsoft lowers its price on Internet Explorer (or lowered its *effective* price for Windows by including Internet explorer in Windows), then more computers could be sold, which means more copies of Windows would be sold *and* more copies of Microsoft's applications – Word, Excel, etc. – would be sold. This means that a lower price for Internet Explorer or Windows could give rise to higher sales, prices, and profits on the applications.

MANAGER'S CORNER: **Covering Relocation Costs of New Hires**

Major corporations are constantly hiring workers from one part of the country only to ask them to move to another part, often a more expensive part. They also often ask their employees to relocate, moving them from one location with a low cost of living to another location with a higher cost of living. Few question whether the firms ordering the movement should pay the cost of the moving van and travel. The trickier issue is whether companies need to fully cover the difference in the cost of living.

As you can imagine, our best answer is that "it depends." But we can do better than that. We can show that if the cost-of-living difference is spread across all goods bought by the relocating workers, the living cost difference will likely have to be covered. However, if the cost difference is concentrated in any one good, for example, housing, the firm can get by with increasing the relocating workers' salaries by less than the cost-of-living difference.

To see these points, which allow us to deduce general principles, suppose that your company's headquarters is in La Jolla, California, where the cost of housing is much higher than in many other parts of the country. Suppose also that you want to hire an engineer from Six Mile, South Carolina where the cost of housing is relatively low. In fact, suppose you learn that the cost of housing in La Jolla is exactly five times the cost of housing in Six Mile. A modestly equipped 2,000 square foot house in La Jolla on a one-tenth of an acre lot, for example, sells for about $500,000. Approximately the same house can be bought in Six Mile (with much more land) for $100,000.

The engineer you are interested in hiring is earning $100,000 a year in Six Mile. In your interviews with the engineer, she tells you, quite honestly, that she likes the job you have for her. However, she also informs you that after comparing La Jolla with her hometown she has found that housing is the only major cost difference. That is, there are minor cost differences for things like food, clothing, and medical care, but those differences wash out, especially after considering quality differences. The two areas are substantially different, she admits, but she values the amenities in the two locations more or less the same. La Jolla has the ocean close by, but Six Mile has the mountains just a short distance to the west.

However, the engineer stresses that at an interest rate of 8.5 percent, the $400,000 additional mortgage she will have to take out to buy a house in La Jolla that is comparable to the one she has back in Six Mile means an added annual housing expenditure for her of $34,000. Therefore, she wants you to compensate her for the difference in the cost of housing, which implies an annual salary of $134,000 (plus she expects all moving and adjustment costs to be covered by your firm).

Do you have to concede to her demands? Many managers do succumb to the temptation to concede to such demands. But, assuming that she is being truthful when she says that the amenities of the areas and the other costs of living balance out, the answer is emphatically, No. You should be able to get by with paying her something less than $134,000 a year. There are two ways of explaining the "no" answer. First, you should recognize that the engineer is getting a lot of purchasing power back in Six Mile in one good, housing. If you gave her the demanded $34,000 in additional salary, she would be able to replace her Six Mile house in La Jolla. However the money payment you provide is fungible, which means that she could buy any number of other things with the added income, including more time at the beach (than she spent in the mountains back in Six Mile) or more meals out (and there are far more restaurants in La Jolla).

Hence, the engineer would actually prefer the $134,000 annual income in La Jolla than the $100,000 income in Six Mile, which goes a long way toward explaining why she is pressing the issue. If that is the case, she could also be happier in La Jolla with something less than $134,000 in salary than she is in Six Mile. To get her to take your job, all you need to do is make her slightly better off at your company's location than she is in Six Mile. Doing that does not require full compensation in the housing cost.

Another way of making the same point, but with greater clarity, is through the use of Figure 8.21, which contains a representation of the engineer's income constraints (or "budget lines" for those who remember their formal economics training) in the two locations. To make the analysis as simple as possible, and stay within the constraints of the two-dimensional graph, we consider two categories of goods: housing, which is on the horizontal axis, and a representative bundle of all other goods on the vertical axis.

The figure shows that with her $100,000 salary in Six Mile, the engineer can buy H_1 units of housing, if she spent all of her income on housing (which, admittedly, would never be practical), or she could buy A_1 bundles of all other goods, if she bought no housing (which is also not practical). More than likely, the engineer will buy some combination of housing and all other goods, say, combination a, H_2 of housing and A_2 of all other goods.

If the engineer were only to get the same $100,000 in income in La Jolla, she would have to choose from the combinations along the inside curve, which extends from A_1 (meaning she could still buy, at the limit, the same number of bundles of all other goods) to H_3 (much less housing if only housing were bought). Clearly, the engineer would be unlikely to take an offer of $100,000, simply because there is no combination along A_1H_3 that is superior to combination a in Six Mile.

If you conceded to her demand of $134,000 in annual income, her income constraint would be the thin line that is parallel to A_1H_3 and goes through a.[18] Clearly, she could be as well off in La Jolla at such a salary because she could still take combination a, but is she likely to do that?

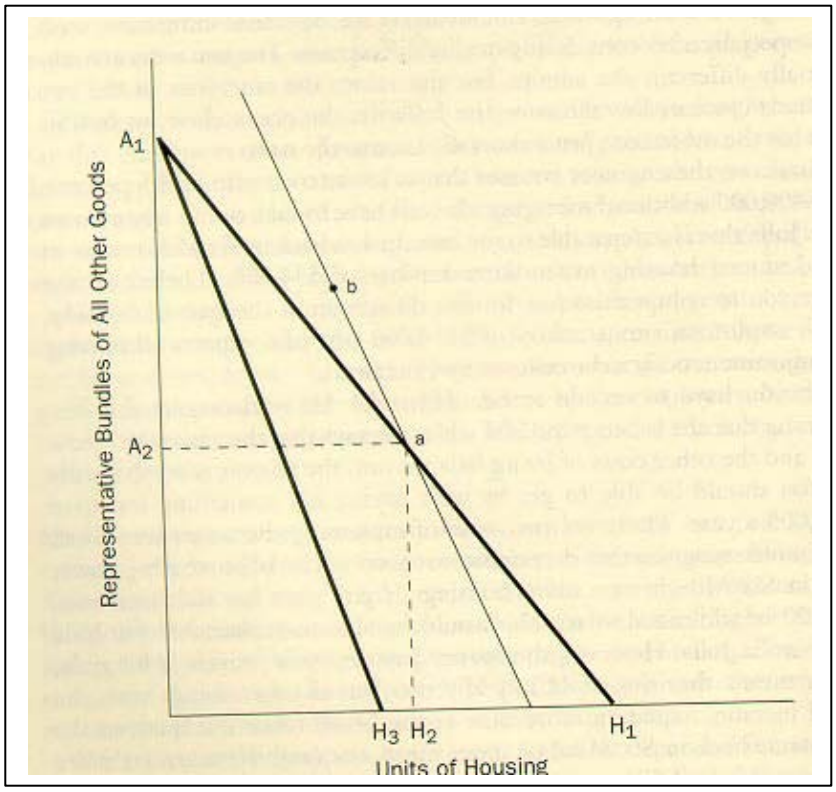


FIGURE 8.21 Choosing between Housing and Bundles of Other Goods

The budget line in Six Mile is A_1H_1 with an income of $100,000. The budget line in L Jolla is A_1H_3 with the same income. If the employer were to offer the engineer a salary of $134,000, which cover the additional cost of housing, the engineers budget line would be the thin line cutting A_1H_1 at a. Hence, the engineer could choose combination b and be better off than in Six Mile. This means that the employer can offer the engineer less than $134,000.

The answer is not likely, because of the changes in relative prices. The price of housing in La Jolla is much higher than the price of housing in Six Mile, which is why her dashed income constraint is much steeper than her old income constraint (A_1H_1).

[18]By giving the engineer $134,000, she can buy the exact combination of goods that she had back in Six Mile, A_2 and H_2. The extra $34,000 in salary would go totally to housing, leaving her with the same amount of after-housing income that she had in Six Mile. Her new income constraint line is parallel with A_1H_3 simply because the prices of the bundles and housing are the same as under A_1H_3, and the relative prices of those goods determine the slope of the income constraint.

The "law of demand" (the economist's analytical pride and joy), which says that price and quantity of goods and services are inversely related, can be expected to apply to housing in our example. Hence, the engineer will likely buy less housing and more of other goods, which implies a movement toward the vertical axis. She very likely will choose a combination like *b*. She will obviously be better off there because were she not, she would have remained with consumption bundle *a*. If she is better off, then you can cut her income below $134,000, taking part of the gains she would otherwise get.

We can't say, theoretically, exactly how little you can pay the engineer. All we can say is that, given the conditions of this problem, you don't have to pay her what she asks, $134,000. You might be able to pay her $130,000 or $125,000 -- something between $100,000 and $134,000. That's not much help, but it is some help, especially given that many of our previous students, when given the problem, think that the engineer's demands would have to be met.

The only time her demands would have to be met is when the added cost of living in La Jolla were distributed more or less evenly among all goods, not just concentrated in housing (which, for those who know both areas of the country, is where a sizable share of the cost differential actually is). This leads us to the conclusion that the more concentrated the cost differential between two areas, the less of the overall cost differential must be made up in the form of salary, or money income, and vice versa.

Of course, this leads to another useful insight. If you are looking for an employee who is living in an area where the cost of living is lower than yours, then you can save on salary by looking where the lower cost of living is concentrated in a single good, such as housing. Conversely, if you are thinking about moving your plant to a "low cost area" like Six Mile, then don't expect to save in salaries an amount that is equal to the difference in the cost of living. You will be able to lower your salaries, but not by the entire cost of living differential.

Of course, we understand that our problem has been relatively simple, given that we have assumed away many of the differences between the two locations. Candidates appraise locations differently. Some people like urban life and the pacific coastal areas, and other people like rural areas and the mountains of the Appalachian region. Those comparative likes will ultimately, of course, go into determining the salary that you will have to pay. You may want someone who is competent to do the job you have, but that is not all that you will be concerned about. You might take someone who is less competent than someone else simply because that person appreciates the amenities of your area more than other more competent candidates, which means that you can get the targeted less-competent person for less. That person may not produce as much, but he or she can still be more cost effective.

When talking about their hiring processes, business people almost always talk about getting the "best" person. We think there is some truth in what they say, but we also know that business people are not always *completely* accurate. What business people should really want is the most *cost-effective* person, and that person is not necessarily, or even often, the most competent.

Our way of looking at the complicated process of business hiring is obviously not fully descriptive of what actually goes on. We can't deal with all the complications here, and would not want to waste your time if we could. We are suggesting, perhaps, some new thoughts, drawn from the economic way of thinking. Our way of looking at the problem also provides guidance in the search for job candidates.

Consider a somewhat different problem. Suppose that you have located two engineers who are candidates for your job, and both live in places like Six Mile that have much lower housing costs than La Jolla. Which one do you choose -- in order to minimize the cost of the new hire to your firm? Of course, you would look at their credentials, but everyone knows to do that. You want the most productive person, but you also want to get the new hire for as little as possible.

Suppose that both candidates are equally productive. What do you do then? If housing is the biggest cost differential, you look at (or ask about) the sizes of their houses, and you should then choose to focus your recruiting efforts on the candidate with the smallest house. Why? You can get that candidate for a lower salary, everything else equal. He or she has a low preference for housing, as revealed by the choice made. The person who has a $100,000 house in Six Mile needs a salary of something less than $134,000 in La Jolla (to compensate for the additional $400,000 mortgage). The candidate who has a $300,000 house in Six Mile (which is likely to be the largest house for miles around) will need a salary of something less than $202,000 (to compensate for the $1.2 million in additional mortgage).

This point can also be made graphically. Consider Figure 8.22, in which lines A_1H_1 and A_1H_3 of Figure 8.21 are replicated. A person who buys combination *b*, including a relatively small house in Six Mile, would require an additional income of something less than the horizontal distance *ab* (which is the additional income that the person needs to duplicate in La Jolla his or her Six Mile house). A person who buys combination *d*, which includes a much larger house in Six Mile, would require an additional income of something less than *cd*. In the graph, *cd* is about twice the size of *ab*.

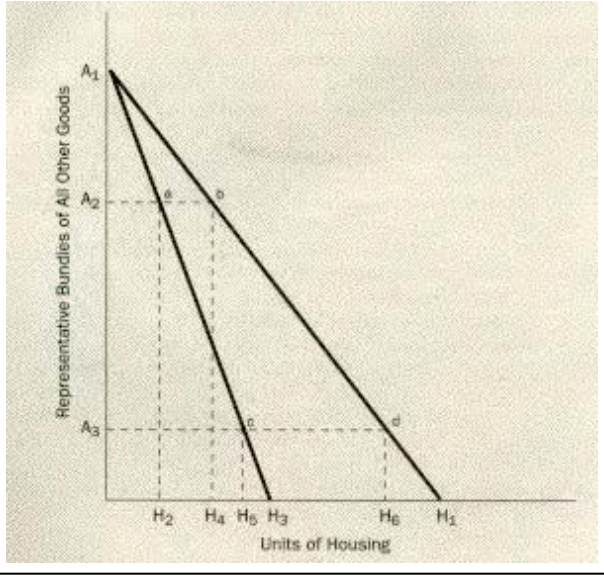


FIGURE 8.22 Choosing Employees Based on the Sizes of their Houses

An employee who chooses combination *b* in Six Mile, with H_4 housing would require additional pay equal to *ab*. An employee who chooses combination *d* in Six Mile would require much more in additional pay, *cd*.

Of course, we recognize that you might -- just might -- be able to find someone with a large house in a place like Six Mile who might take the lower offer. We are only using comparative house sizes as a useful guide for narrowing the search or, in other ways, lowering the cost of your search. The person with a mansion in Six Mile is, in short, likely to be a hard sell.

What you really want to find is someone who has a small house in a place like Six Mile and who is crazy about the beach and the moderate climate near the coast in Southern California. Indeed, one of the often-overlooked reasons for interview trips is not only to assess the person's likely ability on the job, but also to assess how much he or she likes the new location relative to his or her established location. People who like your location relatively more can simply be had for less.

We need to return to the question we started with, Should relocating workers be compensated for housing cost differences? The answer is a qualified no. That is, if housing makes up the main cost difference, then workers moving to a higher housing cost location would be too well compensated if the full cost of living difference were paid. He or she would take less. How much less is a problem that can only be solved by way of interviews and negotiations.

We caution, however, that our analysis flows from an unstated but important assumption, that the housing cost difference in the two locations reflects actual *cost* differences that are not offset by benefit differences. That is, many times, a questionable assumption. Property near the coast in Southern California is much more expensive than in many (but not all) other parts of the country. It is also much more expensive than similar property fifty or a hundred miles inland, but still in California. We must ask why property is so expensive and why so much of the cost difference is in the land that any house sits on. An acre of land in Six Mile may cost no more than a few thousand dollars. On the other hand, an acre in La Jolla (at this writing) can cost upwards of a cool $1 million (a fact that explains why lots are measured in square feet)!

Why the difference? Obviously the demand for property is much higher in La Jolla than in Six Mile, which implies that a lot of people must see some added benefits for being in La Jolla. This implies that for a lot of people, the full difference in housing cost between the two areas need not be covered by added monetary income. A part of the difference in living cost is covered by the "non-money income" associated with the additional amenities in La Jolla that are not compensated for in Six Mile.

The first rule of management (and other disciplines) has sometimes been stated as, "Different Strokes for Different Folks." In our foregoing discussion, we do not mean to suggest that everyone would want to live in La Jolla. If that were the case, the price of land in La Jolla would be far higher than it already is. We mean only to point out that "cost of living" differences cited by business people are not always relevant cost differences because of benefit differences.

To make our point in more concrete terms, it may be true that the measured "cost of living" in La Jolla is 30 percent higher than the cost of living in Six Mile and, for that matter, 30 percent higher than the average for the rest of the country. However, no one should conclude that the cost of *doing business* in La Jolla (or any other "high cost" area)

is 30 percent higher than other parts of the country. The so-called cost of living can be offset in part by amenities and in part by more productive people who are attracted to the high-cost area. Many people with limited productivity will simply not be able to compete with their more productive counterparts in their search for property.[19] In making their employment decisions, firms need to keep these considerations in focus. They need to look carefully at what is implied by "cost of living."

Concluding Comments

Demand is not what people would like to have or are willing to buy at a given price. Rather it is the inverse relationship between price and quantity, a relationship described by a downward sloping curve.

Although economists do not have complete confidence in all applications of the law of demand, they consider the relationship between price and quantity to be so firmly established, both theoretically and empirically, that they call it a law. In difference curves provide a way of "structuring" consumer preferences and deriving the law of demand. In the real world, when the price of a good goes down, the quantity purchased may fall rather than rise. In such cases, economists normally assume (until strong evidence is presented to the contrary) that some other variable has changed, offsetting the positive effects of the reduction in price.

Still, it must be remember that not all downward sloping demand curves are alike. They differ radically in terms of the elasticity of demand, or the responsiveness of consumers to a price change. Managers of public *and* private entities must be aware that the elasticity of demand can affect their business (pricing) strategies.

Review Questions

1. What role does the law of demand play in economic analysis?
2. If the price of jeans rises and the quantity sold goes up, does the demand curve slope upward? Why or why not?

[19]We should, therefore, expect people in high cost areas like La Jolla to have relatively high incomes. One reason is obvious: People need a high-income to cover the high cost of living. Another reason can go unnoticed: People who live in high-cost-of-living areas get much of their income in non-money forms, that is, in the amenities of the area, and these non-money forms of income are not subject to the high marginal tax rates that high-income people pay. For example, people who live on the coast in Southern California have to pay high prices for their housing partly because of the climate, which is very temperate (with high temperatures in the 70s) for much of the year. Accordingly, they have modest heating and cooling bills, which increase the demand and prices of their houses relative to other parts of California and the Southwest where the climate is more extreme and the heating and/or cooling bills are much higher. Of course, pretty scenery can also increase the demand for houses. People in Boulder have been known to say (or lament) that they have to "eat the mountains," meaning their food and household budgets are constrained by the high prices of their houses, inflated by the views of the Rocky Mountains they have.

3. If the prices of most goods are rising by an average of 15 percent per year, but the price of gasoline rises just 10 percent per year, what is happening to the real, or relative, price of gasoline? How do you expect consumers will react?
4. Suppose that a producer raises the price of a good from $4 to $7, and the quantity sold drops from 250 to 200 units. It is demand for the good elastic or inelastic?
5. If the campus police force is expanded and officers are instructed to increase the number of parking tickets they give out, what will happen to the number of parking violations? What may be necessary to eliminate all parking violations? (Why may that option be rejected?)
6. If the government subsidizes flood insurance, what will happen to the price of that insurance? What will happen to the value of the property that is lost during floods? Why?
7. If the price of ballpoint pens falls, will the demand for ballpoint pens change? What will happen to the demand for pencils? To the demand for paper?
8. If a nation appreciates its currency in relation to other national currencies, what will be the effect on other nations' exports and imports? On the willingness of that nation's citizens to invest abroad?
9. Will a tax on imports and a subsidy on exports have the same effect on trade as depreciation of a nation's currency?

PERSPECTIVE: **Experimentally Determined Indifference Curves**

An experiment to determine the characteristics of an individual's indifference curves was performed by K.R. MacCrimmon and M. Toda with seven students from the University of California at Los Angeles. The seven students were asked to construct indifference curves for money and ballpoint pens and for money and pastries. A separate experiment was conducted for each indifference curve. Each experiment began with an initial reference point, or bundle, containing a given amount of money, measured along the horizontal axis, but none of the other good. The student was then presented with bundles containing varying amounts of money and the other good and asked whether each new bundle was preferred or not preferred to the initial bundle. After repeating this a number of times, a rather concise area remained that contained bundles the student found just as attractive as the initial bundle. The student then constructed his or her indifference curve within this area. This experiment was repeated seven times for the money-pen choices and four times for the money-pastry choices, and each experiment was begun with a different amount of money. So each student constructed seven indifference curves for money and pens and four indifference curves for money and pastries.

To motivate students to give thoughtful and honest answers, one of the bundles that had been considered was randomly chosen after each indifference curve was constructed. If it had been preferred to the initial bundle, the student received it; otherwise, the student received the initial bundle containing only money. In the experiments dealing with money and pastries, the student had to eat all the pastries in the bundle received before the money was awarded.

The resulting indifference curves were checked to see if they exhibited the characteristics that economists attribute to indifference curves. The indifference curves for each student were overlaid on the same graph to see if any of them intersected. They did not. The money-pen indifference curves and the

money-pastry indifference curves were non-intersecting for all students. (The money-pastry indifference curves for three students did merge together as they moved out over the money axis.)

Also, as expected, the money-pen indifference curves were downward sloping. Students would give up money only in return for more pens, and vice versa. In other words, both money and pens were considered goods, not bads. This was not true of the money-pastry indifference curves. When the bundles being considered contained only a few pastries, male students would give up a little money to obtain another pastry so that their indifference curves were downward sloping. But after consuming about three pastries, they would consume another pastry only if they received more money. At this point, pastries became a bad, and the indifference curves became upward sloping. For the two women in the experiment, even the first pastry was a bad, and their money-pastry indifference curves were upward sloping from the beginning.

With only one minor exception, the indifference curves were convex everywhere. This is in keeping with the assumption of the normal shapes for indifference curves (which exhibit a diminishing marginal rate of substitution, at least within the vicinity of their tangency with the budget constraint). On the upward-sloping portions of the money-pastry indifference curves, this convexity meant that the more pieces of pastry that were consumed, the more money that would be required to encourage a subject to consume another pastry.

Based on K.R. MacCrimmon and M. Toda, "The Experimental Determination of Indifference Curves," *The Review of Economic Studies* (October 1969), pp. 433-51.

CHAPTER 9

Production Costs and Business Decisions

The economist's stock in trade—his tools—lies in his ability to and proclivity to think about all questions in terms of alternatives. The truth judgment of the moralist, which says that something is either wholly right or wholly wrong, is foreign to him. The win-list, yes-no discussion of politics is not within his purview. He does not recognize the either-or, the all-or-nothing situation as his own. His is not the world of the mutually exclusive. Instead, his is the world of adjustment, of coordinated conflict, of mutual gain.

James M. Buchanan

Cost is pervasive in human action. Managers (as well as everyone else) are constantly forced to make choices, to do one thing and not another. Cost -- or more precisely, opportunity cost -- is the most highly valued opportunity not chosen. Although money is a frequently used measure of cost, it is not cost itself.

Although we may not recognize it, cost also pervades our everyday thought and conversation. When we say "that course is difficult" or "the sermon seemed endless," we are indicating the cost of activities. If the preacher's extended commentary delayed the church picnic, the sermon was costly. Although complaints about excessive costs sometimes indicate an absolute limitation, more often they merely mean that the benefits of the activity are too small to justify the cost. Many people who "can't afford" a vacation actually have the money but do not wish to spend it on travel, and most students who find writing research papers "impossible" are simply not willing to put forth the necessary effort.

This chapter explores the meaning of cost in human behavior. We will begin by showing how seemingly irrational behavior can often be explained by the hidden costs of a choice. We will then develop the concept of marginal cost, which together with demand and the related concept of supply defines the limits of rational behavior, from personal activities like painting and fishing to business decisions like how much to produce.

Inevitably, points made earlier will be reviewed and extended in this chapter. There is a cost in this repetition, but there is also some benefit in a few varied reiterations. We will use the cost analysis to make points that seem to defy common sense in business. For example, we will show that a firm should not necessarily seek to produce at the level at which the average cost of production is minimized.

Explicit and Implicit Costs

Not all costs are obvious. It is not difficult to recognize an out-of-pocket expenditure—the monthly price you pay for a product or service. This is called an explicit cost. **Explicit cost** is the money expenditure required to obtain a resource, product, or service. For example, the price of your book is an explicit cost of taking a course in economics. Other costs are less immediately apparent. Hidden costs of the course might include the time spent going to class and studying, the risk of receiving a failing grade, and the discomfort of being confronted with material that may challenge some of your beliefs. These are implicit costs; together they add up to the value of what you could have done instead. **Implicit cost** is the forgone opportunity to do or squire something else or to put one's resources to another use. Although implicit costs may not be recognized, they are often much larger than the more obvious explicit costs of an action. (Then, there are some "costs" that are recognized on accounting statements that should not be considered in making business decisions. These costs are called "sunk costs." See the box on the next page.)

The Cost of an Education

A good illustration of the magnitude of implicit costs is the cost of an education. Suppose an MBA student—Eileen Payne—takes a course and pays $2,000 for tuition and $200 for books. The money cost of the course is $2,200, but that figure does not include the implicit costs to the student. To take a course, Eileen must attend class for about 45 hours and may have to spend twice that much time traveling to and from class, completing class assignments, and studying for examinations. The total number of hours spent on any one course, then, might be 135 (30 hours in class plus 105 hours of traveling, studying, and so forth).

The student could have spent that time doing other things, including working for a money wage. If Eileen's time is valued at $25 per hour (the wage she might have received if working), the time cost of the course is $3,375 (135 hours x $6). Moreover, if she experiences some anxiety because of taking the course, that psychic or risk cost must be added to the total as well. If Eileen would be willing to pay $500 to avoid the anxiety, the total implicit cost of taking the course climbs to $820.

Explicit costs	
Tuition	$2,000
Books	200
Total explicit cost	$2,200
Implicit costs	
Time	$3,375
Anxiety	500
Total implicit cost	$3,875
Total costs of course	$6,075

The opportunity cost of the student's time represents the largest component of the total cost of the course. The value of one's time varies from person to person. For students who are unable to find work, the time costs of taking a course may be quite small. That is why many young people go to college. Their time cost is generally lower than that of experienced workers who must give up the opportunity to earn a good wage in order to attend classes full time.

PERSPECTIVE: **Why "Sunk Costs" Don't Matter**

*A **sunk cost** is a past cost. Economists define past costs as historical costs that cannot be altered by current decisions. Such costs are beyond the realm of choice. Will a rational, profit-maximizing business firm base its current decisions on its historical costs?*

An example can help to answer this question. Suppose an oil exploration firm purchases the mineral rights to a particular piece of property for $1 million. After several month of drilling, the firm concludes that the land contains no oil (or other valuable mineral resources). Will the firm reason that, having spent $1 million for the mineral rights, it should continue to look for oil on the land? If the chances of finding oil are nonexistent, the rational firm will cease drilling on the land and try somewhere else. The $1 million is a sunk cost that will not influence the decision to continue or cease exploration. Indeed, the firm may begin drilling on land for which it paid far less for mineral rights, if management believes that the chances of finding oil are higher there than on the $1 million property.

The underlying reason that sunk costs do not matter to current production decisions is that in the economist's use of the term, sunk costs are not really costs. The opportunity cost of an activity is the value of the best alternative not chosen. In the case of an historical cost, however, there are no longer any alternatives. Although the oil exploration firm at one time could have chosen an alternative way to spend the $1 million, once the choice was made the alternative ceased to be available. Nor can the firm resell the mineral rights for $1 million; those rights are now worth far less because of accumulated evidence that the land contains little or no valuable minerals. Sunk costs, however painful the memory of them might be, are gone and best forgotten by the firm. Profits are made by looking forward, not backward.

The Cost of Bargains

Every Wednesday, supermarkets run large newspaper ads listing their weekly specials. Generally only a few items are offered at especially low prices, for store managers know that most bargain seekers can be attracted to the store with just a few carefully selected specials. Once the customer has gone to the store offering a special on steak, he would have to incur a travel cost in order to buy other items in a different store. Even though peanut butter may be on sale elsewhere, the sum of the sale price and the travel cost exceed the regular price in the first store. Through attractive displays and packaging, customers can be persuaded to buy many other goods not on sale, particularly toiletries, which tend to bear high markups.

Supermarket chains do not necessarily make huge profits. The grocery industry is reasonably competitive, and supermarket chains as a group are not highly profitable compared to other corporations. The stores manage to recoup some of the revenues lost on sale items by charging higher prices on other goods. In other words, the cost of a bargain on sirloin steak may be a high price for toothpaste.

Some shoppers make the rounds of the grocery stores when sales are announced. For such people, time and transportation are cheap. A person who values his or her time at $10 an hour is not going to spend an hour trying to save a dollar or two. The cost of gas alone can make it prohibitively expensive to visit several stores. Because of the costs of acquiring information, many shoppers do not even bother to look for sales. The expected benefits are simply not great enough to justify the information cost. These shoppers enter the market "rationally ignorant."

Marginal Cost

So far we have been considering cost as the determining factor in the decision to undertake a particular course of action. The rational person weight the cost of an action against it benefits and comes to a decision: whether to invest in an education, to shop around for a bargain, or to operate an airplane. The question is, how much of a given good or service will an individual choose to produce or consume? How does cost limit a behavior once a person has decided to engage in it? The answer lies in the concept of marginal cost.

Rational Behavior and Marginal Cost

Marginal cost is the additional cost incurred by producing one additional unit of a good, activity, or service. Marginal cost is the cost incurred by reading one additional page, making one additional friend, giving one additional gift, or going one additional mile. Depending on the good, activity, or service in question, marginal cost may stay the same or vary as additional units are produced. For example, imagine that Jan smith wants to give Halloween candy to ten of her friends. In a sense, Jan is producing gifts by procuring bags of candy. If she can buy as many bags as she wants at a unit price of fifty cents, the marginal cost of each additional unit she buys is the same, fifty cents. The marginal cost is constant over the range of production.

Marginal cost can vary with the level of output, however, for two reasons. The first has to do with the opportunity cost of time. Suppose Jan wants to give each friend a miniature watercolor, which she will paint herself over the course of the day. To make time for painting, Jan can forgo any of the various activities that usually make up her day. She may choose to give up recreational activities, housekeeping chores, or time spent on work or study.

If she behaves rationally, she will give up the activities she values least. To do the first painting, she may forgo straightening up her room—an activity that is low on most people's lists of preferences. The marginal cost of her first watercolor is therefore a messy room. To paint the second watercolor, Jan will give up the more next-to-last item on her list of favorite activities. As she produces more and more paintings, Jan will forgo more and more valuable alternatives. In other words, the marginal cost of her paintings will rise with her output.

If the marginal cost of each new painting is plotted against the quantity of paintings produced, a curve like the one in Figure 9.1 will result. Because the marginal

cost of each additional painting is higher than the marginal cost of the last one, the curve slopes upward to the right.

Although the marginal cost curve is generally assumed to slope upward, as the one in Figure 9.1 does, that need not be the case. If Jan placed equal value on all the forgone activities, her marginal cost would be constant and the marginal cost curve would be horizontal.

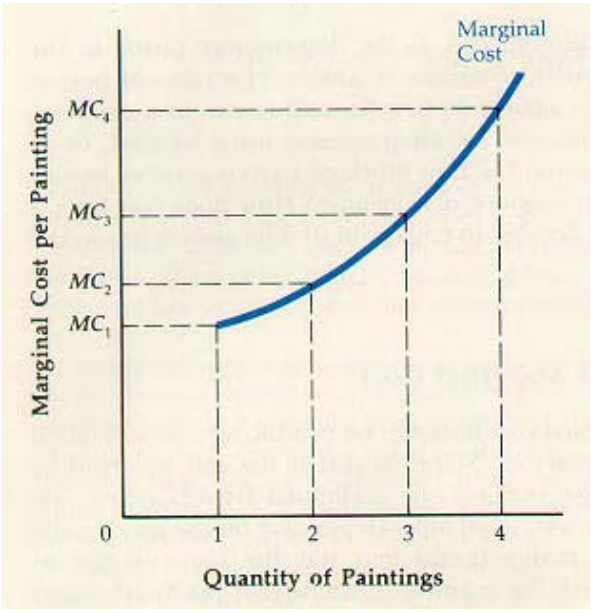


FIGURE 9.1 Rising Marginal Cost
To produce each new watercolor, Jan must give up an opportunity more valuable than the last. Thus the marginal cost of her paintings rises with each new work.

The Law of Diminishing Returns

The second reason marginal cost may vary with output involves a technological relationship known as the **law of diminishing marginal returns**. According to the law of diminishing marginal returns, as more and more units of one resource -- labor, fertilizer, or any other resource -- are applied to a fixed quantity of another resource -- land, for instance -- the increase in total added output gained from each additional unit of the variable resource will eventually begin to diminish. In other words, beyond some point less output is received for each added unit of a resource. That is, more of the resource will be required to produce the same amount of output as before. Beyond some point, the marginal cost of additional units of output rises.

Although the law of diminishing returns applies to any production process, its meaning is most easily grasped in the context of agricultural production. Assume you are producing tomatoes. You have a fixed amount of land (an acre) but can vary the quantity of labor you apply to it. If you try to do planting all by yourself -- dig the holes, pour the water, insert the plants, and core them up -- you will waste time changing tools. If a friend helps you, you can divide the tasks and specialize. Less time will be wasted in changing tools.

The time you would have spent changing tools can be spent planting more tomatoes, thus increasing the harvest. At first, output may expand faster than the labor force. That is, one laborer may be able to plant 100 tomatoes an hour; two working together may be able to plant 250 an hour. Thus the marginal cost of planting the additional 150 plants is lower than the cost of the first 100. Up to a point, the more workers, the greater their efficiency, and the lower the marginal cost—all because of the economies of specialization. At some point, however, the addition of still more laborers will not contribute as much to production as in the past, if only because a large number of workers on a single acre of ground will start bumping into one another. Then the marginal cost of putting plants into the ground will begin to rise.

Diminishing returns are an inescapable fact of life. If returns did not diminish at some point, output would expand indefinitely and the world's food supply could be grown on just one acre of land (For that matter, it could be grown in a flower box.) The point at which output begins to diminish varies from one production process to the next, but eventually all marginal cost curves will slope upward to the right, as in Figure 9.1.

Table 9.1 shows the marginal cost of producing tomatoes with various numbers of workers, assuming that each worker is paid $5 and that production is limited to one acre. Working alone, one worker can produce a quarter of a bushel; two can produce a full bushel (columns 1 and 2). The third column shows the amount each additional worker adds to total production, called the marginal product. **Marginal product** is the increase in total output that results when one additional unit of a resource—for example, labor, fertilizer, and land -- is added to the production process, everything else held constant. The first worker contributed 0.25 (one quarter) of a bushel; the second worker, an additional 0.75 of a bushel, and so on. These are the marginal products of successive units of labor.

The important information is shown in the last two columns of the table. Although two workers are needed to produce the first bushel (column 4), because of the efficiencies of specialization, only one additional worker is needed to produce the second. Beyond that point, however, returns diminish. Each additional worker contributes less, so that two more workers are needed to produce the third bushel and give more to produce the fourth. If the table were extended, each bushel beyond the fourth would require a progressively larger number of workers.

Column 5 shows that if all workers are paid the same wage, $5, the marginal cost of a bushel of tomatoes will decline from $10 for the first bushel to $5 for the second before rising to $10 again for the third bushel. That is, increasing marginal costs (or diminishing returns) emerge after the addition of the third worker.

If the marginal cost of each bushel (column 5) is plotted against the number of bushels harvested, a curve like the one in Figure 9.2 will result. Although the curve slopes downward at first, for most purposes the relevant segment of the curve is the upward-sloping portion above point *a*, will be explained in detail later).

TABLE 9.1 Marginal Costs of Producing Tomatoes

Number of Workers Employed (1)	Total Number of Bushels (2)	Contribution of Each Worker to Production (Marginal Product) (3)	Number of Workers Required to Produce Each Additional Bushel (4)	Marginal Cost of Each Bushel, Figured at $5 per Worker (5)
1	0.25	0.25		
2	1.00	0.75 (1st bushel)	2	$10
3	2.00	1.00 (2nd bushel)	1	$ 5
Point at Which Diminishing Maginal Returns Emerge				
4	2.60	0.60		
5	3.00	0.40 (3rd bushel)	2	$10
6	3.30	0.30		
7	3.55	0.25		
8	3.75	0.20 (4th bushel)	5	$25
9	3.90	0.15		
10	4.00	0.10		

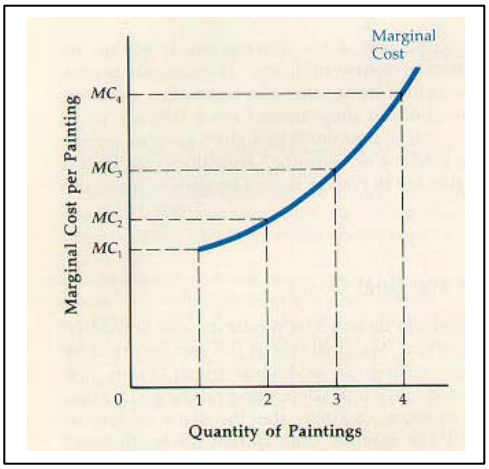


FIGURE 9.2 The Law of Diminishing Marginal Returns

As production expands with the addition of new workers, efficiencies of specialization initially cause marginal cost to fall. At some point, however—here, just beyond two bushels—marginal cost will begin to rise again. At that point, marginal returns will begin to diminish and marginal costs will begin to rise.

The Cost-Benefit Tradeoff

Just as a producer's marginal cost schedule shows the increasing cost of supplying more goods, the demand curve, as explained earlier, shows the decreasing value or marginal benefit of those goods to the people consuming them. Together, marginal costs and benefits determine how many units will be produced and consumed up to the intersection of the marginal cost and demand (marginal benefit) curves, the marginal benefit of each

additional unit exceeds it marginal cost. In other words, people can gain through production and consumption of those units. The intersection of the two curves represents the limit of production, or the point at which welfare is maximized. To see this point, consider the costs and benefits of an activity like fishing.

The Costs and Benefits of Fishing

Gary Schmidt likes to fish. What he does with the fish he catches is of no consequence to us; he can make them into trophies, give them away, or store them in the freezer. Even if Gary places no money value on the fish, we can use dollars to illustrate the marginal costs and benefits of fishing to Gary. (Money figures are not values, but a means of indicating relative value.)

What is important is that Gary wants to fish. How many fish will he catch? From our earlier analysis of Jan's desire to paint (page 181), we know that the cost of catching each additional fish will be higher than the cost of the one before. Gary will confront an upward-sloping marginal cost curve like the one in Figure 9.3. Gary's demand curve for fishing will slope downward, for as the cost of catching each additional fish rises, Gary will be less and less inclined to spend more time on the activity (see Figure 9.3).

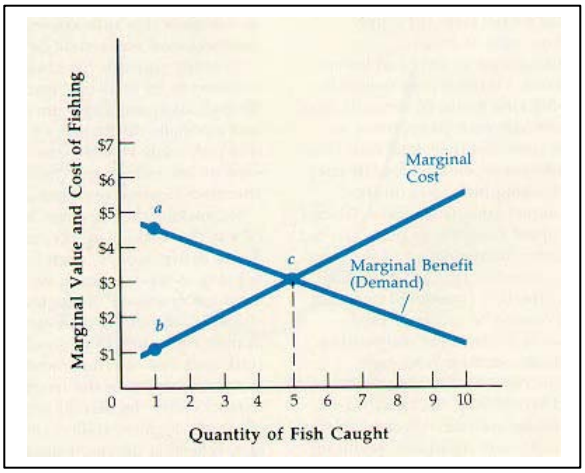


FIGURE 9.3 Costs and Benefits of Fishing

For each fish up to the fifth, Gary receives more in benefits than he pays in costs. The first fish gives him $4.67 in benefits (point *a*) and costs him only $1 (point *b*). The fifth yields equal costs and benefits (point *c*), but the sixth costs more than it is worth. Therefore Gary will catch no more than five fish.

From the positions of the two curves, we can see that Gary will catch up to five fish before he packs up his rod and heads for home. He places a relatively high value of $4.67 on the first fish (point *a* in the figure) and places the relatively low marginal cost of $1 on forgone opportunities for it (point *b*). In other words, he gets $3.67 more value from using his time, energy, and other resources to fish than he wold receive from his nexr best alternative. The marginal benefit of the second fish also exceeds its marginal cost, although by a small amount ($2.75-$4.25 -- $1.50). Gary continues to gain with the third and fourth fishes, but the fifth fish is a matter of indifference to him. Its marginal value equals its marginal cost (point *c*). Although we cannot say that Gary will actually bother to catch a fifth fish, we do know that five is the limit toward which he will aim.

He will not catch a sixth—at least during the period of time offered by the graph—because it would cost him more than he would receive in benefits.

The Costs and Benefits of Preventing Accidents

All of us would prefer to avoid accidents. In that sense we have a demand for accident prevention, whose curve should slope downward like all other demand curves. Preventing accidents also entails costs, however, whether in time, forgone opportunities, or money. Should we attempt to prevent all accidents? Not if the cost of ensuring that you will never stumble down the stairs is $100 (again, we are using dollars to indicate relative value). If the only injury you expect to suffer were a bruised knee, would you spend $100 to prevent the accident?

As with the question of how long to fish, marginal cost and benefit curves can help illustrate the point at which preventing accidents ceases to be cost effective. Suppose Al Rosa's experience indicates that he can expect to have ten accidents over the course of the year. If he tries to prevent all of them, the value of preventing he last one, as indicated by the demand curve in Figure 9.4, will be only $1 (point *a*). The marginal cost of preventing it will be much greater: approximately $6 (point *b*)**.** If Al is rational, he will not try to prevent the last accident. As a matter of fact, he will try to prevent only five accidents (point *c*). As with the tenth accident, it will cost more than it is worth to Al to prevent the sixth through ninth accidents. He would try to prevent all ten accidents only if his demand for accident prevention were so great that his demand curve intersected the marginal cost curve at point *b*.

Some accidents may be unavoidable. In that case, the marginal cost curve will eventually become vertical. Other accidents may be avoidable in the sense that it is physically possible to take measures to prevent them—although the rational course may be to allow them to happen.

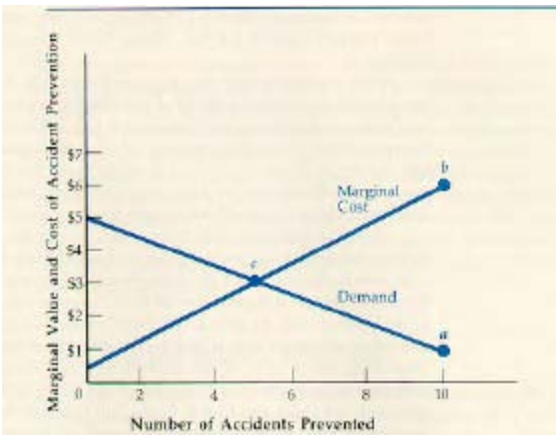


FIGURE 9.4 Accident Prevention

Given the increasing marginal cost of preventing accidents and the decreasing marginal value of preventing the accidents, *c* accidents will be prevented.

The Production Function in Pictures

Business firms combine various factors of production in order to produce various goods and services. Although there are thousands of different factors of production, or inputs,

for simplicity we often use a model with only two factors, labor and capital. We can then study how the two inputs can be combined to produce an output. The relationship between inputs and output is called the **production function**. The general equation for the production function is:

$$Q = f(L, K)$$

where Q is output, L is labor, K is capital, and f is the functional relationship between inputs and output. In the short run, we assume that capital cannot be varied; labor is therefore, the only variable factor. To increase output, then, a firm must increase the amount of labor.

The relationship between the amount of the variable input (labor) and output can be illustrated with a total product curve such as that in the upper half of Figure 9.5. Suppose that the curve is that of a commercial fishing firm. The firm's capital—the boat and equipment—is fixed in the short run. Only the number of workers can vary. As the amount of labor increases from zero, the fish catch (output) increases. Between zero and 5 workers, output increases at an increasing rate. As more workers are hired total output continues to increase, although at a decreasing rate, until 15 workers are hired. Beyond that point, hiring more workers *reduces* output.

The reason the total product curve has that particular shape can be seen more clearly in the lower half of Figure 9.5, which shows the average and marginal product curves. The average product of labor is total output divided by the amount of labor, or Q/L. The marginal product of labor is the change in total output brought about by changing the amount of labor by one unit. Because at least some workers are needed to operate the boat and the equipment, the first few workers hired greatly increase total output; marginal product is rising. Between 5 and 15 workers, the marginal product of labor falls, although the average product continues to rise (because it is less than marginal product). Total product continues to rise, but no longer at an increasing rate. The law of diminishing marginal returns has taken effect. At seven workers, marginal product equals average product and average product is maximized. As more workers are hired average product falls. Note that as long as marginal product is positive, more labor means more output and the total product curve will have a positive slope. Beyond 15 workers, marginal product becomes negative and total product falls. The boat may be so crowded that workers bump into each other and reduce the amount of work that each does. To catch more fish once this stage has been reached, the firm must buy a larger boat.

Some economists divide the production function of Figure 9.5 into three stages. In stage one, from zero to seven workers, total product and average product of labor both rise. In stage two, between seven and 15 workers, total product rises while average product falls. In stage three, beyond 15 workers, total product and average product both fall (and marginal product is negative).

Price and Marginal Cost: Producing to Maximize Profits

"Production" is not generally an end in itself in business. Most firms seek to make a profit. How can we think about how they go about the task of trying to maximize profits? The total and marginal product curves need to be converted to cost curves. Only then can we engage in familiar cost-benefit analyses.

Granted, many business people derive intrinsic reward from their work. They may value the satisfaction of producing a product that meets a human need just as much as the profits they earn. Some business people may even accept lower profits so their products can sell at lower prices and serve more people. For most business people, however, the profit generated by sales is the major motivation for doing business.

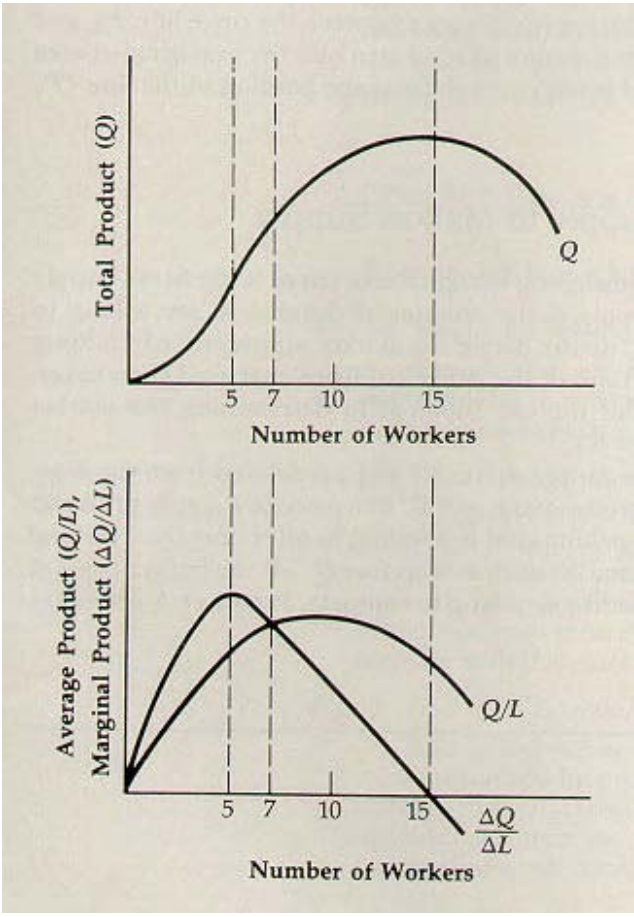


FIGURE 9.5 Total, Average, and Marginal Product Curves

The total product curve shows how output changes when the amount of the variable input, labor, changes. Total product rises first at an increasing rate (0 to 5 workers), then at a decreasing rate (5 to 15 workers), before declining (beyond 15 workers). The marginal and average product curves reflect what is happening to total product. Marginal product rises when total product is rising at an increasing rate and falls when total product is rising at a decreasing rate. Marginal product is positive when total product is rising and negative when total product is falling.

How much will a profit-maximizing firm produce? Assume its marginal cost curve is like the one in Figure 9.6(a). Assume further that the owners can sell as many units as they want at a price of P_1. Because this firm is in business to make a profit, the price of its product can be thought of as the marginal benefit of each additional unit. P_1 is also the firm's marginal revenue. **Marginal revenue** is the additional revenue a firm

acquires by selling an additional unit of output. Each time the firm sells one additional unit, its revenues rise by P_1.

Clearly, a profit-maximizing firm will produce and sell any unit for which the marginal revenue acquired (*MR*) exceeds the marginal cost (*MC*). (Profits are the difference between total costs and total revenues. Therefore a firm's profits rise whenever an increase in revenues exceeds the increase in its costs.) At a price of P_1, then, this firm will produce up to, and no more than, Q_1, products. For every unit up to Q_1, price is greater than marginal cost.

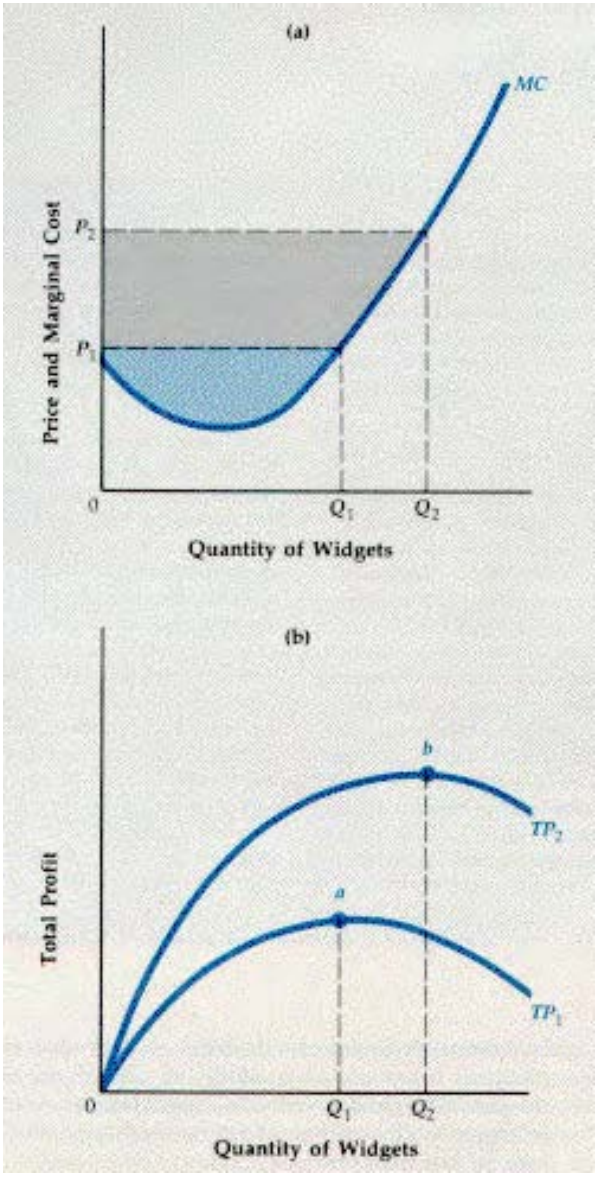


FIGURE 9.6 Marginal Costs and Maximization of Profit

At price *P1* (part (a)), this firm's marginal revenue, shown by the shaded area under P1, exceeds its marginal cost up to an output level of *Q1*. At that point total profit, shown in part (b), peaks (point *a*). At price *P2*, marginal revenue exceeds marginal cost up to an output level of Q2. The increase in price shifts the profit curve in part *b* upward, from *TP1* to *TP2*, and profits peak at *b*.

The vertical distance between P1 and the marginal cost of each unit, as shown by the marginal cost curve, is the additional profit obtained from each additional unit produced. By summing the vertical distance between P1 and the marginal cost curve for all units up to Q_1, we can obtain the firm's total profits. (See the color-shaded area in Figure 9.5(a).) Total profits can also be represented as a curve, as in the line TP_1 in Figure 9.5(b). Notice that the curve peaks at Q_1 the point at which the firm chooses to stop producing. Beyond Q_1, marginal cost is greater than marginal revenue, and total profits fall, as shown by the downward slope of the total profits curve.

What will the firm do if the price of its product rises from P_1 to P_2? For the firm that can sell all it wants at a constant price, a rise in price means a rise in marginal revenue. Once the price rises to P_2, the marginal revenue of an additional $Q_2 - Q_1$ products exceeds their marginal cost. At the higher price, a larger number of units can be profitably produced and sold. The firm will seek to produce up to the point at which marginal cost equals the new, higher marginal revenue, P_2, or output, Q_2, in Figure 9.5 (a). As before, profit is equal to the vertical distance between the rice line, P_2, and the marginal cost curve, or the color-shaded area plus the gray-shaded area in Figure 9.5 (a). The total profit curve shifts to the position of the line TP_2 in Figure 9.5(b).

From Individual Supply to Market Supply

If a portion of the upward-sloping marginal cost curve is the firm's supply curve, and if market supply is the amount all producers are willing to produce at various prices, we can obtain the market supply curve by adding together the elevation portions of the individual firms' marginal cost curves. (This procedure resembles the one followed in determining the market demand curve in an earlier chapter.)

Figure 9.7 shows the supply curves S_A and S_B, derived from the marginal cost curves of two producers, A and B. At a price of P_1, only producer B is willing to produce anything, and it is willing to offer only Q_1. The total quantity supplied to the market at P_1 is therefore Q_1. At the higher prices of P_2, however, both producers are willing to compete. Producer A offers Q_1, while producer B offers more, Q_2. The total quantity supplied is therefore Q_3 the sum of Q_1 and Q_2.

The market supply curve, S_{A+B} is obtained by adding the amounts A and B are willing to sell at each price and splitting the totals. Note that the market supply curve lies farther from the origin and is flatter than the individual producers' supply curves. The entry of more producers will shift the market supply curve farther outward and lower its slope even more. (More will be said about cost and supply in later chapters.)

MANAGER'S CORNER: **Cutting Health Insurance Costs**

The cost of doing business is a constant worry for all firms. At times, those business costs feed major policy debates in the nation's capital. As is so often the case, the infamous "healthcare crisis" in the United States amounts to nothing more than costs for

a particular service – healthcare -- responding to the market forces of supply and demand. Unfortunately, the forces have been distorted by legal and political factors that have gotten the incentives wrong. In our view, the "crisis" is more a matter of political rhetoric than economics. Political grandstanding alone will hardly solve whatever healthcare problem exists. Careful reflection by policy makers and managers on the exact sources of the problem might. The current distortion presents a possibility for managers to benefit both their firms and its workers by policies that get the incentives right.

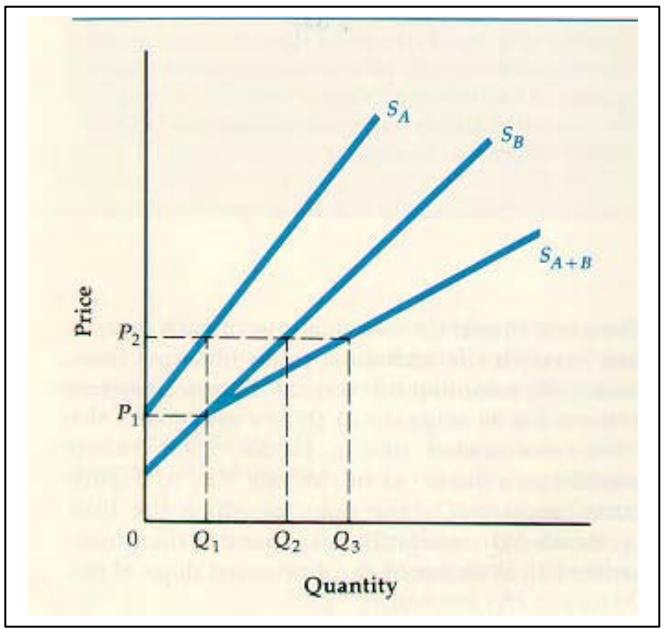


FIGURE 9.7 Market Supply Curve

The market supply curve (S_{A+B}) is obtained by adding together the amount producers A and B are willing to offer each at each and every price, as shown by the individual supply curves S_A and S_B. (The individual supply curves are obtained from the upward sloping portions of the firms' marginal cost curve.)

If private firms and Washington-based politicians want to reform the system and temper cost increases, they can do so by working with the forces of supply and demand, which means, fundamentally, changing people's incentives to provide and consume healthcare services.

Granted, healthcare costs, and the insurance premiums that finance a major share of healthcare expenditures, have risen faster than the prices of other goods over the last couple of decades.[1] Indeed, the cost of health insurance provided by firms was escalating at double-digit rates in the late 1980s and the very early 1990s when increases in the consumer price index, a broad measure of the cost of living, were falling.[2] In the mid-1990s, healthcare cost increases slowed, but they were, at this writing, still increasing at a rate that was over 50 percent higher than the rate of increase in the general cost of living.

[1] See Paul J. Feldstein. Health Policy Issues: An Economic Perspective on Health Reform (Arlington, Va. : AUPHA Press; Ann Arbor, Mich.: Health Administration Press, 1994); and Paul J. Feldstein, The Politics of Health Legislation: An Economic Perspective, 2nd ed. (Chicago, Ill.: Health Administration Press, 1996).

[2] Put another way, the consumer price index was increasing at decreasing rates, which means that the rate of inflation was gradually but irregularly decreasing for most of the 1980s and 1990s.

In order to understand the problem of insurance cost increases, we need first to consider the market forces that have been at work driving up healthcare costs. What are those forces? Consider the following list of factors affecting the supply and demand of healthcare:

1. Doctors have been subject to a growing degree of litigation. They have been sued with growing frequency partly because they have made mistakes, but also because they are now being held responsible for problems over which they may have no control. Patients have found that they can make money by blaming doctors for almost any problem that emerges when they are being treated. Fearful that they will be sued for delivering incomplete or misguided care, doctors have been covering their financial and professional backsides by ordering tests that may be only marginally valuable from a medical perspective but can help them defend themselves in the event they are sued when problems emerge. They have also been trying to acquire legal protection and to spread the risk of lawsuits by increasing referrals to specialists.
2. Federal expenditures on Medicare for older patients and Medicaid for low-income patients have increased the demand for healthcare services since the late 1960s, which has tended to boost prices and forced many younger and lower-income patients out of the health insurance market.
3. Medical care has become technologically more sophisticated, and doctors have applied the new technology for offensive reasons (to keep patients alive longer) and for defensive reasons (they don't want to be accused of negligence for failing to employ the latest life-saving technology). The extensive use of the latest and best technology may have saved and prolonged lives, but medical care costs have been driven up in the process.
4. The healthcare industry has always been plagued by the problem of "asymmetric information," or the doctors knowing more about many patients' medical conditions and what will remedy their problems than do the patients themselves. As a consequence, doctors have always been in a position to induce patients to buy more medical care than the patients might really buy, if they had the information and knowledge at the disposal of the doctors.
5. Medical technology has drastically lowered the cost of many medical procedures and has, as a consequence, lowered the cost of extending the lives of patients by some varying and uncertain number of months and years. For example, less than four decades ago, heart and kidney transplants and heart bypass operations were impossible. No one knew how to do them. Then, the costs of those procedures were infinite. Their prices may now remain high in absolute dollar terms, running into the tens, if not hundreds, of thousands of dollars. However, those high prices also represent *lower* prices. And the lower prices for those procedures have, no doubt, increased the number of patients who have been willing and able to pay for the procedures (as well as insurers who have helped with the payments). Although the issue has not been statistically evaluated to date, the lower prices for many medical procedures have probably increased total medical expenditures in

absolute dollar terms and as a percentage of national income. Hence, some of the so-called healthcare "crisis" probably mirrors, to a degree, the success of the healthcare industry in lowering the cost of prolonging life.

6. The cost of employer-provided medical insurance is tax deductible, which means that its price has been artificially lowered, causing more consumers to buy more complete insurance coverage and to demand more medical services (than they otherwise would). The greater demand has enabled medical professionals to boost their prices. As tax rates rose in the 1960s and 1970s, workers naturally had growing incentive to take more of their income in tax-deductible fringe benefits and less of an incentive to take their income in taxable money wages. The higher tax rates spurred demand for health insurance and healthcare – and added to pressure on healthcare costs.

7. Employers have typically bought insurance policies with very low deductibles, for example, $200 a year. This means that after the first $200 of medical care expenditures in any one year, the cost of additional medical services to the insured patient is often close to zero. This feature of insurance policies has encouraged excessive use of healthcare services, which, in turn, has driven up employees' insurance premiums and caused some workers to forgo health insurance altogether.[3]

8. The growth in social problems -- crimes involving bodily injury, the use of street drugs, and teenage pregnancy -- has also contributed to the demand for medical services, which has driven up their prices as well as the price of insurance. The unwillingness or inability of medical professionals to deny services to people who cannot pay for the services has also increased the number of people seeking services. Social attitudes favoring universal medical care coverage have reduced the cost of irresponsible behavior, increasing the demand on the healthcare industry and inflating costs.

Without question, if the grocery industry were operated the way the healthcare industry operates, then we would likely have a "crisis" in the grocery business. The reason is simple: People would pay a fixed sum each month (their grocery premium) through their employer that would entitle them to virtually unlimited access to the grocery store shelves (after they have covered the $200 annual deductible) at zero, or very low, cost. Under such an arrangement, we should not be surprised if people consumed significantly more and better food, some of which would have limited value. We should also not be surprised if the shoppers' grocery premiums went through the roof as everyone allowed their tastes to run wild, with many low-income shoppers forced out of the grocery policies by the inflated premiums.

---

[3] As you may recall from our study of consumer behavior in the last chapter, a working rule of consumer maximizing behavior is that the consumer will continue to buy units of any good or service until the point at which the marginal cost of the last unit consumed just equals the marginal value of the last unit. If the person consumes more than that amount, the additional cost of any additional units will exceed their additional value. By "excessive" consumption, we mean that patients are induced to go beyond the point where the marginal value is, while still positive, less than the marginal cost. The reason for this excessive consumption is that the individual consumer isn't paying the entire cost of additional medical care.

How can the so-called "crisis" be solved, at least partially? We don't intend to offer a detailed set of public policy solutions here. Other specialists in the field have done that.[4] We only point out here that many of the supply and demand forces listed above are beyond the control of individual businesses. There is simply not much most individual businesses can do to affect the broad sweep of social attitudes and government tax and expenditure policies. We only note, however, that the demand for healthcare services can be lowered by reducing, at least marginally, government subsidies for the healthcare of many Americans. This can be accomplished by lowering Medicare and Medicaid expenditures and by eliminating all or a part of the tax deductibility of health insurance. The cost of healthcare can also be lowered by reducing the rewards from suing doctors or by giving patients the right (to a greater or lesser degree) to absolve doctors of liability for problems that they may encounter while the patients are in the doctors' care.

Frankly, making those recommendations is much easier than getting them passed. They are too politically painful for voters (although we suggest that voters should also consider the gains to everyone from getting healthcare costs under control).

Barring changes in public policies, what can businesses themselves do to ameliorate their own healthcare costs? Many businesses have done what has come naturally: they have tried to select workers who are not likely to have medical problems and, therefore, drive up the firms' insurance costs. This is, we remind you, a solution that can benefit both owners *and* many workers, given that healthier workers can mean lower labor costs for firms and lower health insurance premiums. While people might object to this solution on fairness grounds, we stress that it is the type of discriminatory hiring policy that is likely to emerge when health insurance costs have been distorted by political factors, such as the ones included in the list above.

Another private policy solution can emerge if employers and employees recognize that low deductibles on health insurance policies are very expensive because they encourage workers to spend someone else's money, which motivates excessive demand for healthcare and high insurance premiums. With a deductible of $5,000, the price of an additional dollar of insurance coverage for a forty-year old male is measured as a tiny fraction of a cent (actually, .06 of a cent). However, when the deductible is $500, the price escalates to 55 cents. When the deductible is as low as $100, the price of an additional dollar of coverage rises to $2.14, a poor bargain for owners and their employees.[5]

There is an obvious solution to the health insurance problem that has the potential of not only introducing greater efficiency into the healthcare business but also improving the fairness of the system, without any policy change in Washington. This solution seeks to lower the private demand for healthcare by changing the incentives a firm's workers have to consume healthcare services.

---

[4]See John C. Goodman and Gerald L. Musgrave, Patient Power: Solving America's Health Care Crisis (Washington, D.C. : Cato Institute, 1992).

[5]As reported by Goodman and Musgrave (Ibid.).

As we indicated above, most firms that offer their workers health insurance provide "Cadillac policies," ones with small deductibles and broad coverage for just about everything that can go wrong with a person, regardless of whether the person is responsible, through destructive behaviors, for the problems encountered. Each worker has little incentive not to use healthcare services for the slightest problem. Each worker has less incentive to incur the costs that might be required to eliminate or reduce their destructive behaviors.

Each worker can reason that if he or she were to cut back on personal usage of this or that healthcare service, the company's health insurance costs would not be materially affected. Certainly, the individual's health insurance premiums would not fall by the full value of the healthcare services not utilized. The savings from non-use by any one individual, if the savings are detectable at all, will be spread over the entire group of workers through slightly lower premiums for everyone. In short, the individual gains precious little from personal restraint in consumption of healthcare services.[6] Hence, the individual has little incentive to curb consumption.

Granted, if everyone in a firm were to cut back on healthcare usage, then everyone could possibly gain in terms of reduced insurance premiums. The amount of savings could be substantial, and everyone would share in the savings of everyone else. However, as is so often true in business and, for that matter, all group settings, getting everyone to do what is in their best collective interest comes up against the prisoners' dilemma discussed earlier. If everyone else cuts back, there is still no necessary and compelling reason for any one person to cut back. The one person's reduction is, again, inconsequential -- regardless of what all others do. And, we must add, as we have throughout the book, the larger the group, the more difficult the problem in bringing about collective cohesiveness of purpose.[7]

The basic problem for the firm should be seen as one of finding a means of giving all workers an incentive to cut their consumption. This can be done by raising the price of healthcare usage. But how can the price of healthcare be raised by the firm?

Economist John Goodman, head of the National Center for Policy Analysis, recommends what appears to us to be a ingenious and practical solution, one that firms can, as some already have, institute on their own -- to the benefit of the workers *and* the firm.

To see how Goodman's proposal might work, let us start with a few observations and assumptions. Many firms spend upwards of $4,500 annually per worker on health insurance, partly because, with the small deductible, workers have an incentive to consume a lot of healthcare. Let us assume that a basic *catastrophic* health insurance policy, one with a very large deductible of about $3,000 (meaning the insurance covers

---

[6]Of course, the extent to which the individual's actions can be detected depends on the size of the employment group. In small groups of workers, it would be easier to detect the impact of what one individual does or does not do.

[7]One of the more serious problems in having government provide health insurance is that the relevant *group* is really large, extending to the boundaries of the country, which means people may have absolutely no incentives to curb their consumption of healthcare services. The benefits of doing so are spread ever so thinly over too many people.

only major medical problems), can be purchased for each employee for a premium of $1,200 per year (which is, we are told, in the ballpark of the actual cost for a group policy).

Suppose also that the employer agrees to provide this catastrophic insurance policy and, at the same time, agrees to place in a bank reserve account (what Goodman prefers to call a "Medical Savings Account" or "MSA") a sum of $3,000 each year per employee. The employer tells the employees that they can draw on that account for any medical "need" (with "need" being defined broadly). The workers can use the account, for example, to pay for visits to doctors, to cover the cost of hospital stays not covered by insurance, or to pay for a membership in a fitness center (given that exercise can prevent the need for some medical care). Finally, suppose that the workers are also told that the balance remaining in the account at the end of the year can be applied to their individual retirement accounts, or even withdrawn at the end of the year for any purpose that the workers choose.[8]

This proposal has a chance of lowering the employees' healthcare consumption because it requires that people pay for most routine medical care with their own money. Under common insurance arrangements, the additional cost of medical procedures (other than the patients' time) approximates zero (after the low deductible is met). Under the MSA proposal, the cost to the employee of the first $3,000 of medical care is exactly equal to the cost of the service. This is because the employee is made the *residual claimant* on the balance at the end of the year. Hence, we should expect that workers will more carefully evaluate their usage of medical services and cut back. After all, under the old system, the workers were probably consuming "too much," given the low cost (close to zero) that they incurred.

We would expect that the gains from this new MSA system could be shared by both the workers and their firm. We have already developed the example in a way that obviously benefits the firm. The firm was paying $4,500 a year for the insurance of each worker. Now, it must pay $1,200 for the insurance and $3,000 for the MSA, for a total of $4,200. The firm saves $300 per worker.

The workers, however, can also gain. Under the old arrangement, the workers were getting "paid" with insurance, not money. Under the MSA system, they are given a pot of money, $3,000, that they can use, if they choose, to buy insurance that would cover the first $3,000 of care. But many would not likely do that. They can self-insure just by holding onto the money and paying the first $3,000 in medical bills. However, they can, conceivably, also buy a variety of other things, from new televisions to education programs to additional days of vacation.[9] Accordingly, the additional money should enable workers to be better off by allocating the sum to higher valued uses.

---

[8]The particulars of the Medical Savings Accounts are not important here. The important characteristic is broad discretion on the part of the worker, which will likely mean that the worker has a sum of money that is set aside to cover the large deductible under a catastrophic medical insurance policy and that can be used by the employee when it is not spent for medical purposes.

[9]Any actual MSA program might for political reasons have restrictions on the range of goods and services that the workers can buy with any MSA balance remaining at the end of the year. For example, one MSA-type proposal would require that the balance go into a worker's retirement account.

Both workers and their employers can also gain because the new insurance arrangement can be expected to lower the worker's demand for use of the health insurance provided by their employers. Many workers will want to be careful not to use up their $3,000 account, as they become more careful shoppers of medical care. Workers will make use of the catastrophic insurance only in those situations when they have serious problems and little choice but to make use of medical care, which explains why the premiums for catastrophic insurance are so low.

By providing catastrophic health insurance coupled with a medical savings account, a firm can attract better workers by providing them with a more valuable compensation package at lower cost. Overall, we would expect the firms that adopt this type of insurance system would be more productive and competitive.

However, we hasten to add that our simple example does not reflect the full complexity of employment conditions most firms face. The problem managers will have in developing acceptance of the MSA is the cross-subsidies that are embedded in current insurance programs. Low-risk workers typically subsidize high-risk workers. Hence, we doubt that the firm's deposit into workers' MSA accounts would equal the insurance deductible, as we have assumed in our example. The reason is that many healthy (typically younger) workers are fortunate in that they often don't go to the doctor or hospital in any given year, and other workers have only modest medical expenditures in most years. They are subsidizing the unhealthy (typically older) workers who make extensive use of medical care. If the MSA deposit equaled the deductible, this cross-subsidy would be wiped out, and the insurance company would very likely be hit with high bills from the high-risk workers without the payments from the low-risk workers. To make the MSA system work, the deposit would have to be limited, with the workers themselves sharing in some of the gains in the event they have limited expenses but also sharing in some of the risks if their expenses exceed their MSA deposits. Therein lies the rub, which will rule out many firms from instituting the deal. However, some firms will still be able to find a reasonable compromise.

Managers must also be mindful of the possibility that MSAs can set up perverse incentives for some workers for some types of healthcare. Knowing that they will have to draw down their MSA account in order to cover annual physical examinations (and other preventive healthcare measures), workers can reason that MSAs increase the immediate cost of physical examinations. But that doesn't mean that the "cost" of physicals goes up for all workers. For some cost will rise; for others the cost will fall. Some employees, no doubt, will be more inclined to get physicals, given that physicals can be paying propositions (or will have a lower *net* cost to them). That is to say, the employees can reason that the current outlay from their MSA for a physical can be more than offset by the reduction in MSA outlays in the future, given that current physicals can "nip" health problems when they are minor. Thus, current physicals can lower the workers' healthcare expenditures from their MSA account over the long run.

However, we suspect that it's also a safe bet that some employees will not be able, or will not be willing, to make the required careful calculations or can properly assess the current and future benefits of physicals. Other workers may reason that most of their later healthcare expenditures for "major" problems that go undetected will be

covered as the catastrophic health insurance kicks in. To accommodate these potential problems, employers can consider covering a portion of the current cost of physicals and other preventive measures. The employers can cover the added cost of subsidizing the physicals and preventive care with any reduction in their insurance premiums they get from encouraging preventive care. If there are no insurance savings from the subsidy, then it seems reasonable to conclude that either the problem of employees skipping preventive care is not a problem or it is such a minor problem that the insurance companies see no need to reduce the insurance premiums of firms that encourage preventive care.

The main point is that managers must be tread carefully in trying to accommodate problems with "preventive care." The problem is that "preventive care" can include not only physicals, but also an array of tests that have little useful medical value. If "preventive care" is defined too broadly and the subsidies are high, managers can be back in the prisoner's dilemma trap that results in excessive healthcare and healthcare insurance expenditures, the net effect of which is healthcare benefits that are not worth the costs to the workers.

Has the MSA concept been tried and has it worked? Yes, on both counts, although the trials to date do not correspond exactly with our example above. One of the problems is that Medical Savings Accounts are not tax deductible, which means that a part of the added cost that must be overridden with benefits is the greater tax payments workers and firms must pay. Nevertheless, several firms have already tried the system with beneficial effects:

- After Quaker Oats put $300 in each worker's Medical Saving Account, the company's healthcare costs grew 6.3 percent a year. However, this was during a period when the healthcare costs of the rest of the country were growing at double-digit rates.
- Forbes magazine encourages its employees to curb medical care expenditures with a variation of the MSA, by paying workers $2 for every $1 of medical costs not incurred up to $1,000. This means that if a Forbes employee incurs medical costs of only $300 in a given year, the employee is rewarded with a check of $1,400 at the end of the year [2 x ($1,000 - $300)]. The magazine's healthcare costs fell 17 percent in 1992 and 12 percent in 1993, years during which other firms' insurance costs were rising.
- The utility holding company Dominion Resources gives each worker who chooses a $3,000 deductible on the company's health insurance policy a deposit of $1,650 a year. Since 1989, its insurance premiums have not risen, while the insurance premiums of other companies have risen by an average of 13 percent a year.
- Golden Rule Insurance Company gives each worker a $2,000 deposit if they select a deductible of $3,000. In 1993, its health insurance costs were 40 percent lower than they would have otherwise been.[10]

[10]See "Answering the Critics of Medical Savings Accounts," Brief Analysis (NCPA, September 16, 1994), p. 1.

We don't propose to tell firms what to do in their own particular circumstances for a very good reason: Frankly, we obviously don't know the details of the individual circumstances of what we hope will be a multitude of business readers of this book. We can use our incentive-based approach to explore the *types* of business policies managers should consider and then adjust to fit the particulars of their circumstances. Moreover, our focus on health insurance is only illustrative of insights that are relevant across a firm's entire fringe benefit package.

The important point of this discussion is by now an old one for this book: Incentives matter. One of the several important reasons many workers pay high health insurance premiums is that they don't have much of an incentive to carefully evaluate their healthcare purchases. The best way of ensuring that workers get the most out of their healthcare benefits is one that is as old as business itself: make the buyer pay a price that reflects the true cost of their decision.

Medical Savings Accounts are simply a means (perhaps one of many that have not yet been devised) of making workers potentially better off by making everyone pay a price for what they consume. This solution may not work for all businesses. Some worker groups may not want to be bothered with considering the costs of their behaviors. However, it appears that many firms and their workers have not considered policies like Medical Savings Accounts because they have not realized that they harbor the potential of making everyone better off. These are the types of policies all managers should examine. Such policies can raise their workers' welfare, their firm's stock prices, and the compensation of managers. Again, we return to what is by now an old point of the book: firms can make money not only by selling more of their product or service, but also by creatively restructuring incentives in mutually beneficial ways.

Concluding Comments

Cost plays a pivotal role in a producer's choices. Costs change with the quantity produced. The pattern of those changes determines the limit of a producer's activity—from the production of salable goods and services to the employment of leisure time. The individual will produce a good or service, or engage in an activity, until marginal cost equals marginal benefit (marginal revenue). Graphically, this is the point where the supply and demand curves for the individual's behavior intersect. At this point, although additional benefits might be obtained by producing additional units of the good, service, or activity, the additional costs that would be incurred discourage further production.

Costs will not affect an individual's behavior unless he or she perceives them as costs. For this reason the economist looks for hidden, implicit costs in all choices. Such costs, if uncovered, will affect choices that remain to be made. Implicit costs can also be helpful in explaining those choices that have already been made.

Review Questions

1. Evaluate the adages "haste makes waste" and "a stitch in time saves nine" from an economic point of view.

2. If executives' time is as valuable as they claim, why are they frequently found reading the advertisements in airline magazines en route to a business meeting?
3. The price of a one-minute long distance call on a cell phone is several times the cost of a call on any other phone. Does that mean that the introduction of cell phones has increased the cost of long distance calling?
4. In discussing accident prevention, we assumed an increasing marginal cost. Suppose instead that the marginal cost of preventing accidents remains constant. How will that assumption affect the analysis?
5. Using the analysis of accident prevention, develop an analysis of pollution control. Using demand and supply curves for clean air, determine the efficient level of pollution control.
6. People take some measures to avoid becoming victims of crime. Can the probability of becoming a victim be reduced to (virtually) zero? If so, why don't people eliminate that probability? What does the underlying logic of your answer suggest about the cost of committing crimes and the crime rate?
7. If the money price of a good rises from $5 to $10, the economist can confidently predict that less will be purchased. One cannot be equally confident that denying a child a dessert will improve the child's behavior, however. Explain why.
8. Consider the information in the production schedule that follows. (a) At what output level do diminishing returns set in? (b) Assume that each worker receives $8. Fill in the marginal product column, and develop a marginal cost schedule and a marginal cost curve for the production process.

Number of Workers	Total Product of All Workers	Marginal Product of Each Worker
1	0.10	
2	0.30	
3	0.60	
4	1.00	
5	1.45	
6	2.00	
7	2.50	
8	2.80	
9	3.00	
10	3.19	
11	3.37	
12	3.54	
13	3.70	
14	3.85	
15	4.00	
16	3.90	
17	3.70	

READING: Sunk Costs in the Railroad Industry

Clinton H. Whitehurst, Jr., Clemson University

Historically, a large part of a railroad's investment has been in assets with fixed costs—cost that do not vary with output in the short run. In the early 1900s, fixed costs were estimated to be as much as 75 percent of railroads' total costs. More recently they have been estimated at 40 to 50 percent.

A significant part of a railroad's fixed costs is the investment in its right of way—the 75- to 200-foot-wide corridors in which its tracks are laid. Most railroads purchased that land and paid for its grading many years ago, perhaps in the last century. Those costs are considered historical, or sunk.

To the degree that its costs are fixed, a railroad's average total cost decreases as its volume increases. The more tons it carries per mile, the lower the average total cost of moving a ton of freight. The railroad's fixed costs are simply spread out over more units of freight.

To use their hauling capacity fully and lower their average total cost, railroads have tended to set their rates low for long hauls. In the early days they often generated only enough revenues to cover their variable costs, not their total costs. But in many instances they compensated for low rates on long hauls by charging high rates on short hauls. In 1887, customer complaints about differences in rates prompted congress to place railroad rates and routes under the regulation of the Interstate Commerce Commission (ICC). Throughout much of its history, the ICC considered rates that did not cover total costs to be unfair or predatory—designed, that is, to drive out competition. It insisted that railroads set their rates high enough to cover total costs.

After the Second World War, the rapidly growing trucking industry became the railroads' chief competitor. Fixed costs were much less significant in trucking than in railroads. As much a 90 percent of the total cost of trucking varied with the number of tons carried per mile. From the point of view of the trucking industry, then, the ICC's requirement that rates cover total costs made sense. But from the railroads' perspective, the requirement was disastrous. By keeping railroad rates high, the ICC enabled the trucking industry to compete for railroad business and expand its share of the transportation market.

In 1958, following an extensive lobbying effort by the railroads, Congress amended the Interstate Commerce Act. The amendment instructed the ICC that "Rates of a carrier shall not be held up to a particular level to protect the traffic of any other mode of transportation." Earlier Interstate Commerce Act provisions still barred "unfair or destructive competitive practice," however. Given the ambiguity of the legislation, the ICC continued to insist that rates cover total costs. In 1968 the Supreme Court upheld its interpretation.

Recently railroads have been given considerable freedom to set their own rates under the railroad Revitalization and Regulatory Reform Act (1976) and especially the Staggers Rail Act (1980). Rates that cover only variable costs are no longer considered unfair and are not challenged by the Interstate Commerce Commission.

Meanwhile, the interstate highways—the right of way for trucks—are becoming more congested. Truck delivery, once much faster than railroad delivery, is slowing down. But railroad tracks remain underutilized. As circumstances change, railroads are putting their century-old investment in their rights of way to good use. By ignoring sunk costs and offering lower rates, they have recaptured much of the freight business they lost to trucks after the Second World War. Today, one often sees highway trailers riding on railroad flatcars, reflecting the new competitiveness of railroads. In fact, hauling trailers is now one of the fastest-growing railroad services.

CHAPTER 10

Production Costs in the Short Run and Long Run

In economics, the cost of an event is the highest-valued opportunity necessarily forsaken. The usefulness of the concept of cost is a logical implication of choice among available options. Only if no alternatives were possible or if amounts of all resources were available beyond everyone's desires, so that all goods were free, would the concepts of cost and of choice be irrelevant.

Armen Alchian

The individual firm plays a critical role both in theory and in the real world. It straddles two basic economic institutions: the markets for resources (labor, capital, and land) and the markets for goods and services (everything from trucks to truffles). The firm must be able to identify what people want to buy, at what price, and to organize the great variety of available resources into an efficient production process. It must sell its product at a price that covers the cost of its resources, yet allows it to compete with other firms. Moreover, it must accomplish those objectives while competing firms are seeking to meet the same goals.

How does the firm do all this? Clearly firms do not all operate in exactly the same way. They differ in organizational structure and in management style, in the resources they use and in the products they sell. This chapter cannot possibly cover the great diversity of business management techniques. Rather, our purpose is to develop the broad principles that guide the production decisions of most firms.

Like individuals, firms are beset by the necessity of choice, which as Armen Alchian reminds us, implies a cost. Costs are obstacles to choice; they restrict us in what we do. Thus a firm's cost structure (the way cost varies with production) determines the profitability of its production decisions, both in the short run and in the long run. Of course, there is one very good reason MBA students should know something about a firm's *cost structure*. "Firms" don't do anything on their own. It's really managers who activate firms and make decisions that will ultimately determine whether a firm is profitable or not.

Out analysis of a firm's "cost structure" is nothing like the imagined costs on accounting statements. Accounting statements indicate the costs that were incurred when the firm produced the output that it did. Here, in this chapter, we want to devise a way of structuring costs for many different output levels. The reason is simple: We want to use this structure to help us think through the question of which among many output levels will enable the firm to maximize profits.

You will also notice that our cost structure is very *abstract*, meaning that it is independent of the experience of any given real-world firm in any given real-world industry. We develop the cost structure in abstract terms for another good reason: MBA students plan to work in a variety of industries and in a variety of firms within those different industries. We want to devise a cost structure that is potentially useful in many different business contexts. To do this, we need to construct costs in several different ways for different time periods, because production costs depend critically on the amount of time for production.

Fixed, Variable, and Total Costs in the Short Run

Time is required to produce any good or service. Therefore, any output level must be founded on some recognized period of time. Even more important, the costs a firm incurs vary over time. In thinking about costs, then, we must identify clearly the period of time over which they apply. For reasons that will become apparent as we progress, economists speak of costs in terms of the extent to which they can be varied, rather than the number of months or years required to pay them off. Although in the long run all costs can be varied, in the short run firms have less control over costs.

The **short run** is the period during which one or more resources (and thus one or more costs of production) cannot be changed—either increased or decreased. Short-run costs can be either fixed or variable. A **fixed cost** is any cost that (in total) does not vary with the level of output. Fixed costs include overhead expenditures that extend over a period of months or years: insurance premiums, leasing and rental payments, land and equipment purchases, and interest on loans. Total fixed costs (TFC) remain the same whether the firm's factories are standing idle or producing at capacity. As long as the firm faces even one fixed cost, it is operating in the short run.

A **variable cost** is any cost that changes with the level of output. Variable costs include wages (workers can be hired or laid off on relatively short notice), material, utilities, and office supplies. Total variable costs (*TVC*) increase with the level of output.

Together, total fixed and total variable costs equal total cost. Total cost (*TC*) is the sum of fixed costs and variable costs at each output level.

$$TC = TFC + TVC$$

Columns 1 through 4 of Table 10.1 show fixed, variable, and total costs at various production levels. Total fixed costs are constant at $100 for all output levels (see column 2). Total variable costs increase gradually, from $30 to $395, as output expands from 1 to 12 widgets. Total cost, the sum of all fixed and variable costs at each output level (obtained by adding columns 2 and 3 horizontally), increases gradually as well.

Graphically, total fixed cost can be represented by a horizontal line, as in Figure 10.1. The total cost curve starts at the same point as the total fixed cost curve (because total cost must at least equal fixed cost) and rises from that point. The vertical distance between the total cost and the total fixed cost curves shows the total variable cost at each level of production.

Table 10.1 Total, Marginal, and Average Cost of Production

Production Level (number of widgets) (1)	Total Fixed Costs (2)	Total Variable Costs (3)	Total Costs (2) + (3) (4)	Marginal Cost (change in 3 or 4) (5)	Average Fixed Cost (2) div (1) (6)	Average Variable Cost (3) div (1) (7)	Average Total Cost (4) div (1) or (6) + (7) (8)
1	$100	$ 30	$ 130	$30	$100.00	$30.00	$130.00
2	100	50	150	20	50.00	25.00	75.00
3	100	60	160	10	33.33	20.00	53.33
4	100	65	165	5	25.00	16.25	41.25
5	100	75	175	10	20.00	15.00	35.00
6	100	90	190	15	16.67	15.00	31.67
7	100	110	210	20	14.29	15.71	30.00
8	100	140	240	30	12.50	17.50	30.00
9	100	180	280	40	11.11	20.00	31.11
10	100	230	330	50	10.00	23.00	33.00
11	100	300	400	70	9.09	27.27	36.36
12	100	395	495	95	8.33	32.92	41.25

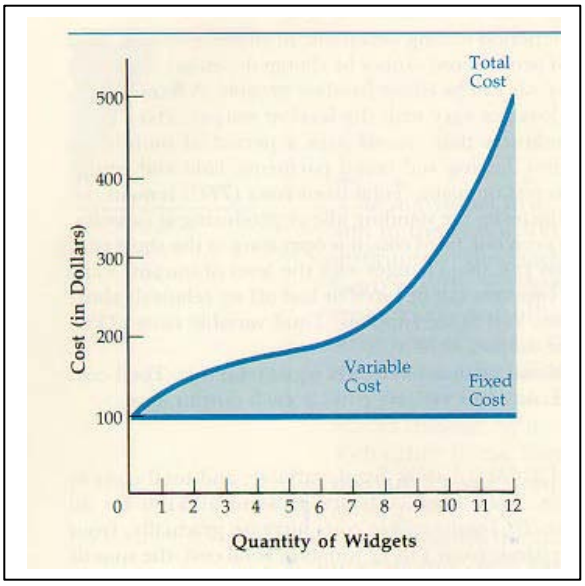


Figure 10.1 Total Fixed Costs, Total Variable Costs, and Total Costs in the Short Run

Total fixed cost does not vary with production; therefore, it is drawn as a horizontal line. Total variable cost does rise with production. Here it is represented by the shaded area between the total cost and total fixed cost curves.

Marginal and Average Costs in the Short Run

The central issue of this and following chapters is how to determine the profit-maximizing level of production. In other words, we want to know what output the firm that is interested in maximizing profits will choose to produce. Although fixed, variable, and total costs are important measures, they are not very useful in determining the firm's

profit-maximizing (or loss-minimizing) output. To arrive at that figure, as well as to estimate profits or losses, we need four additional measures of cost: (1) marginal, (2) average fixed, (3) average variable, and (4) average total. When graphed, those four measures represent the firm's cost structure. A cost structure is the way various measures of cost (total cost, total variable cost, and so forth) vary with the production level. These four cost measures cover all costs associated with production, including risk cost and opportunity cost.

Marginal Cost

We have defined marginal cost (*MC*) as the additional cost of producing one additional unit. By extension, marginal cost can also be defined as the change in total cost. Because the change in total cost is due solely to the change in variable cost, marginal cost can also be defined as the change in total variable cost per unit:

$$MC = \frac{\text{change in TC}}{\text{change in quantity}} = \frac{\text{change in TVC}}{\text{change in quantity}}$$

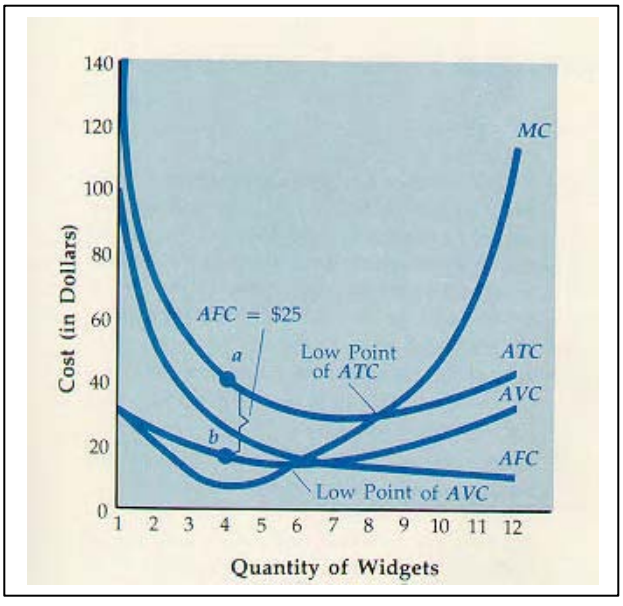


Figure 10.2 Marginal and Average Costs in the Short Run

The average fixed cost curve (*AFC*) slopes downward and approaches, but never touches, the horizontal axis. The average variable cost curve (*AVC*) is mathematically related to the marginal cost curve and intersects with the marginal cost curve (*MC*) at its lowest point. The vertical distance between the average total cost curve (ATC) and the average variable cost curve equals the average fixed cost at any given output level. There is no relationship between the *MC* and *AFC* curves.

As you can see from Table 10.1, marginal cost declines as output expands from one to four widgets and then rises, as predicted by the law of diminishing returns. This increasing marginal cost reflects the diminishing marginal productivity of extra workers and other variable resources the firm must employ in order to expand output beyond four widgets.

The marginal cost curve is shown in Figure 10.2. The bottom of the curve (four units) is the point at which marginal returns begin to diminish.

Average Fixed Cost

Average fixed cost (*AFC*) is total fixed cost divided by the number of units produced *(Q):*

$$AFC = \frac{TFC}{Q}$$

In Table 10.1, total fixed costs are constant at $100. As output expands, therefore, the average fixed cost per unit must decline. (That is what business people mean when they talk about "spreading the overhead." As production expands, the average fixed cost declines.)

In Figure 10.2, the average fixed cost curve slopes downward to the right, approaching but never touching the horizontal axis. That is because average fixed cost is a ratio, *TFC/Q*, and a ratio can never be reduced to zero. No matter how large the denominator (*Q*). Note that this is a principle of arithmetic, not economics.)

Average Variable Cost

Average variable cost is total variable cost divided by the number of units produced, or

$$AVC = \frac{TVC}{Q}$$

At an output level of one unit, average variable cost necessarily equals marginal cost. Beyond the first unit, marginal and average variable cost diverge, although they are mathematically related. Whenever marginal cost declines, as it does initially in Figure 10.2, average variable cost must also decline. The lower marginal value pulls the average value down. A basket ball player who scores progressively fewer points in each successive game for instance, will find her average score falling, although not as rapidly as her marginal score.

Beyond the point of diminishing returns, marginal cost rises, but average variable cost continues to fall for a time (see Figure 10.2). As long as marginal cost is below the average variable cost, average variable cost must continue to decline. The two curves meet at an output level of six widgets. Beyond that point, the average variable cost curve must rise because the average value will be pulled up by the greater marginal value. (After a game in which she scores more points than her previous average, for instance, the basketball player's average score must rise.) The point at which the marginal cost and average variable cost curves intersect is therefore the low point of the average variable cost curve. Before that intersection, average variable cost must fall. After it, average variable cost must rise. For the same reason, the intersection of the marginal cost curve and the average total cost curve must be the low point of the average total cost curve (see Figure 10.2)

Average Total Cost

Average total cost *(ATC)* is total of all fixed and variable costs divided by the number of units produced (*Q*), or

$$ATC = \frac{TFC + TVC}{Q} = \frac{TC}{Q}$$

Average total cost can also be found by summing the average fixed and average variable costs, if they are known *(ATC = AFTC + AVC)*. Graphically the average total cost curve is the vertical summation of the average fixed and average variable cost curves (see Figure 10.2).

Because average total cost is the sum of average fixed and variable costs, the average fixed cost can be obtained by subtracting average variable from average total cost: *AFC = ATC – AVC*. On a graph, average fixed cost is the vertical distance between the average total cost curve and the average variable cost curve. For instance, in Figure 10.2, at an output level of four widgets, the average fixed cost is the vertical distance *ab*, or $25 ($41.25 - $16.25, or column 8 minus column 7 in Table 10.1).

From this point on, the average fixed cost curve will not be shown on a graph, for it complicates the presentation without adding new information. Average fixed cost will be indicated by the vertical distance between the average total and average variable cost curves at any given output.

Marginal and Average Costs in the Long Run

So far our discussion has been restricted to time periods during which at least one resource is fixed. That assumption underlies the concept of fixed cost. Fortunately, over the long run all resources that are used in production can be changed. The **long run** is the period during which all resources (and thus all costs of production) can be changed—either increased or decreased. By definition, there are no fixed costs in the long run. All long-run costs are variable.

The foregoing analysis is still useful in analyzing a firm's long-run cost structure. In the long run, the average total cost curve (ATC in Figure 10.2) represents one possible scale of operation, with one given quantity of plant and equipment (in Table 10.1, $100 worth). A change in plant and equipment, which are no longer fixed, will change the firm's cost structure, increasing or decreasing its productive capacity.

How do changes in long-run costs affect a profit-maximizing firm's production decisions? Generally, they can encourage firms to produce on a larger scale.

Economies of Scale

Figure 10.3 illustrates the long-run production choices facing a typical firm. The curve labeled ATC_1 is, in reduced form, the average total cost curve developed in Figure 10.2. Any additional plant and equipment will add to total fixed costs, and at low output levels (up to q_1) will lead to higher average total costs (curve ATC_2). On the new scale of operation, however, average total cost need not remain high. At higher output levels (q_1 to q_2), the firm may realize economies of scale, cost decreases that stem from an expanded use of resources (see page 29).

Economies of scale can occur for several reasons. Expanded operation generally permits greater specialization of resources. Technologically advanced equipment, like mainframe computers, can be used, and more highly skilled workers can be employed. Expansion may also permit improvements in organization, like assembly-line production. As a firm increases its scale of operation, indivisibility or unavoidable excess capacity of resources declines. The important point is that by spreading the higher cost of additional plant and equipment over a larger output level, the firm can reduce the average cost of production.

Economies of scale cannot necessarily be realized in every kind of production: there are few or no economies of scale in the production of original works of art. The principle will hold true for most production operations, however. Curve ATC_2 in Figure 10.3 cuts curve ATC_1 and then dips down to a lower minimum average total cost—at a higher output level. Curve ATC_3 does the same with respect to curve ATC_2.

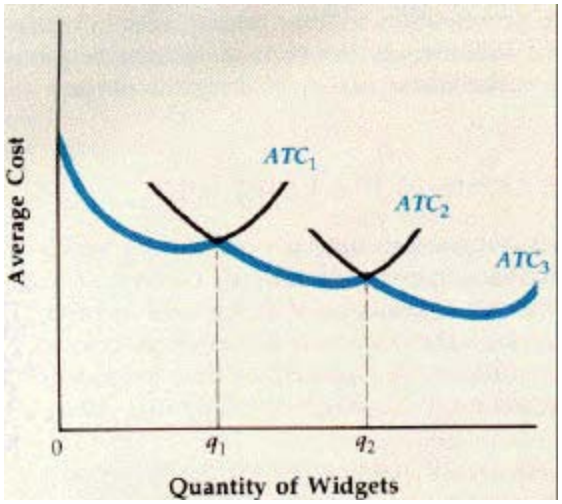


FIGURE 10.3 Economies of Scale

Economies of scale are cost savings associated with the expanded use of resources. To realize such savings, however, a firm must expand its output. Here the firm can lower its costs by expanding production from q_1 to q_2—a scale of operation that places it on a lower short-run average total cost curve (ATC_2 instead of ATC_1).

Diseconomies of Scale

Economies of scale do not last forever. That is to say, a firm cannot increase its use of resources indefinitely and expect its average total cost to continue to fall. At some point, a firm will confront diseconomies of scale—cost increases that stem from an expanded

use of resources.[1] Diseconomies of scale are illustrated in Figure 10.4. Beyond curve ATC_4, an increase in the scale of operation leads to a higher minimum average cost.

Average and Marginal Costs

When will a firm change its scale of operation? In markets filled with risk and uncertainty about actual costs and demand, that is a tough question. Ideally, the firm will change scale as soon as it becomes profitable—in Figure 10.3, at output level q_1. Before q_1 the average cost on scale ATC_1 is lower than the average cost on scale ATC_2. The fixed costs of additional plant and equipment simply cannot be spread over enough output to reduce the average total cost. Beyond q_1, however, the average cost on scale ATC_2 is lower than the average cost on scale ATC_1. Therefore the firm can minimize its overall cost of operation by expanding along the colored portion of the curve ATC_2, and it can push its average costs down even further by expanding its scale once again at output level q_2.

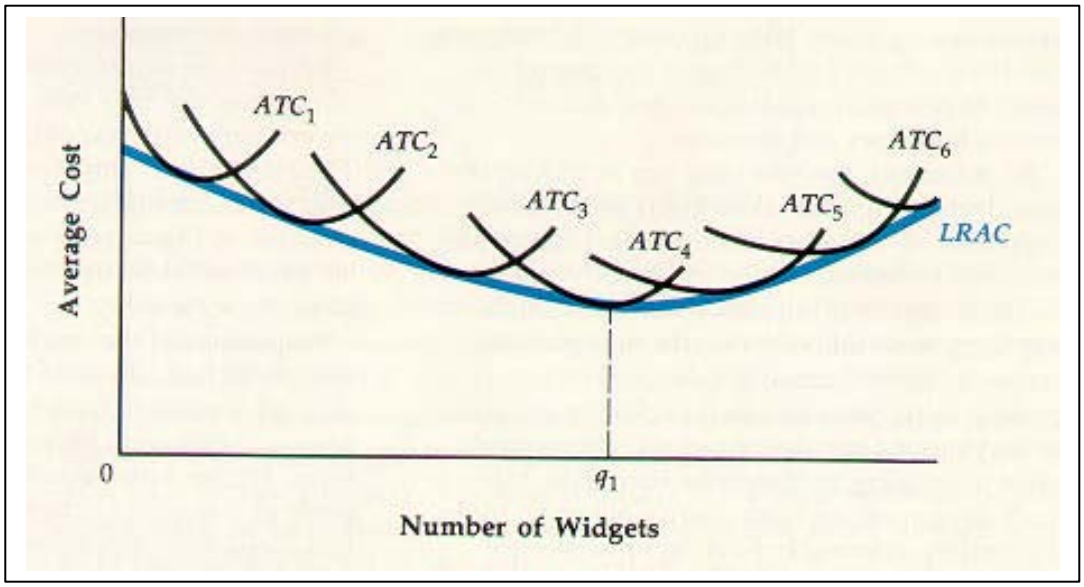


FIGURE 10.4 Diseconomies of Scale

Diseconomies of scale may occur because of the communication problems of larger firms. Here the firm realizes economies of scale through its first four short-run average total cost curves. The long-run average cost curve begins to turn up at an output level of q_1, beyond which diseconomies of scale set in.

[1] For a while, a firm may be able to avoid diseconomies of scale by increasing the number of its plants. Management's ability to supervise a growing number of plants is limited, however, and eventually diseconomies of scale will emerge at the level of the firm, if not the plant. If diseconomies of scale did not exist, in the long run each industry would have only one firm.

Assuming there are many more scales of operation than are represented in Figure 10.3, the firm's expansion path can be seen as a single overall curve that envelops all of its short-run average cost curves. Such a curve is shown in Figure 10.4 and reproduced in Figure 10.5 as the long-run average cost curve (LRAC).

Like short-run average cost curves, the long-run average cost curve has an accompanying long-run marginal cost curve. If long-run average cost is falling, as it does initially in Figure 10.5, it must be because long-run marginal cost is pulling it down. If long-run cost is rising, as it does eventually in Figure 10.5, then long-run marginal cost must be pulling it up. Hence at some point like q_1 long-run marginal cost must turn upward, intersecting the long-run average cost curve at its lowest point, q_2.

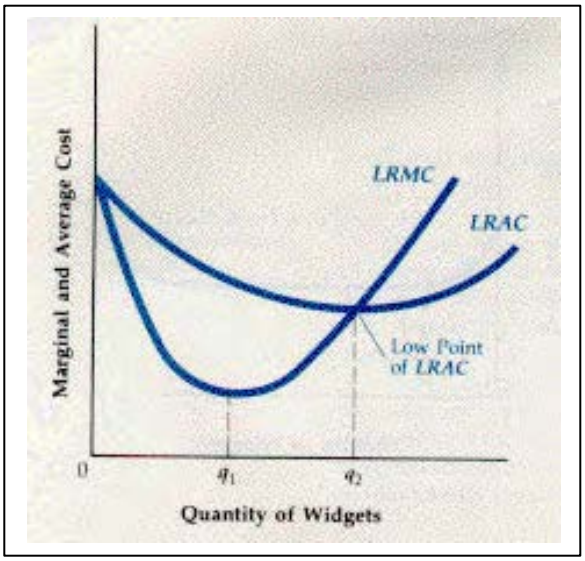


FIGURE 10.5 Marginal and Average Cost in the Long Run

The long-run marginal and average cost curves are mathematically related. The long-run average cost curve slopes downward as long as it is above the long-run marginal cost curve. The two curves intersect at the low point of the long-run average cost curve.

Individual Differences in Average Cost

Not all firms experience economies and diseconomies of scale to the same degree, or at the same levels of production. Their long-run average cost curves, in other words, look very different. Figure 10.6 shows several possible shapes for long-run average cost curves. The curve in Figure 10.6(a) belongs to a firm in an industry with few economies of scale and significant diseconomies at relatively low output levels. (This curve might belong to a firm in a service industry, like shoe repair.) We would not expect profit-maximizing firms in this industry to be very large, for firms with an output level beyond q_1 can easily be underpriced by smaller, lower-cost firms.

Figure 10.6(b) shows the long-run average cost curve for a firm in an industry with modest economies of scale at low output levels and no diseconomies of scale until a fairly high output level. In such an industry—perhaps apparel manufacturing—we would expect to find firms of various sizes, some small and some large. As long as firms are producing between q_1 and q_2, larger firms do not have a cost advantage over smaller firms.

Figure 10.6(c) illustrates the average costs for a firm in an industry that enjoys extensive economies of scale—for example, an electric power company. No matter how far this firm expends, the long-run average cost curve continues to fall. Diseconomies of scale may exist, but if so they occur at output levels beyond the effective market for the firm's product. This type of industry tends toward a single seller—a natural monopoly. A **natural monopoly** is an industry in which long-run marginal and average costs generally decline with increases in production, so that a single firm dominates production. Given the industry's cost structure, that is, one firm can expand its scale, lower its cost of operation, and underprice other firms that attempt to produce on a smaller, higher-cost scale. Electric utilities have been thought for a long time to be natural monopolies (which has supposedly justified their regulation, a subject to which we will return).

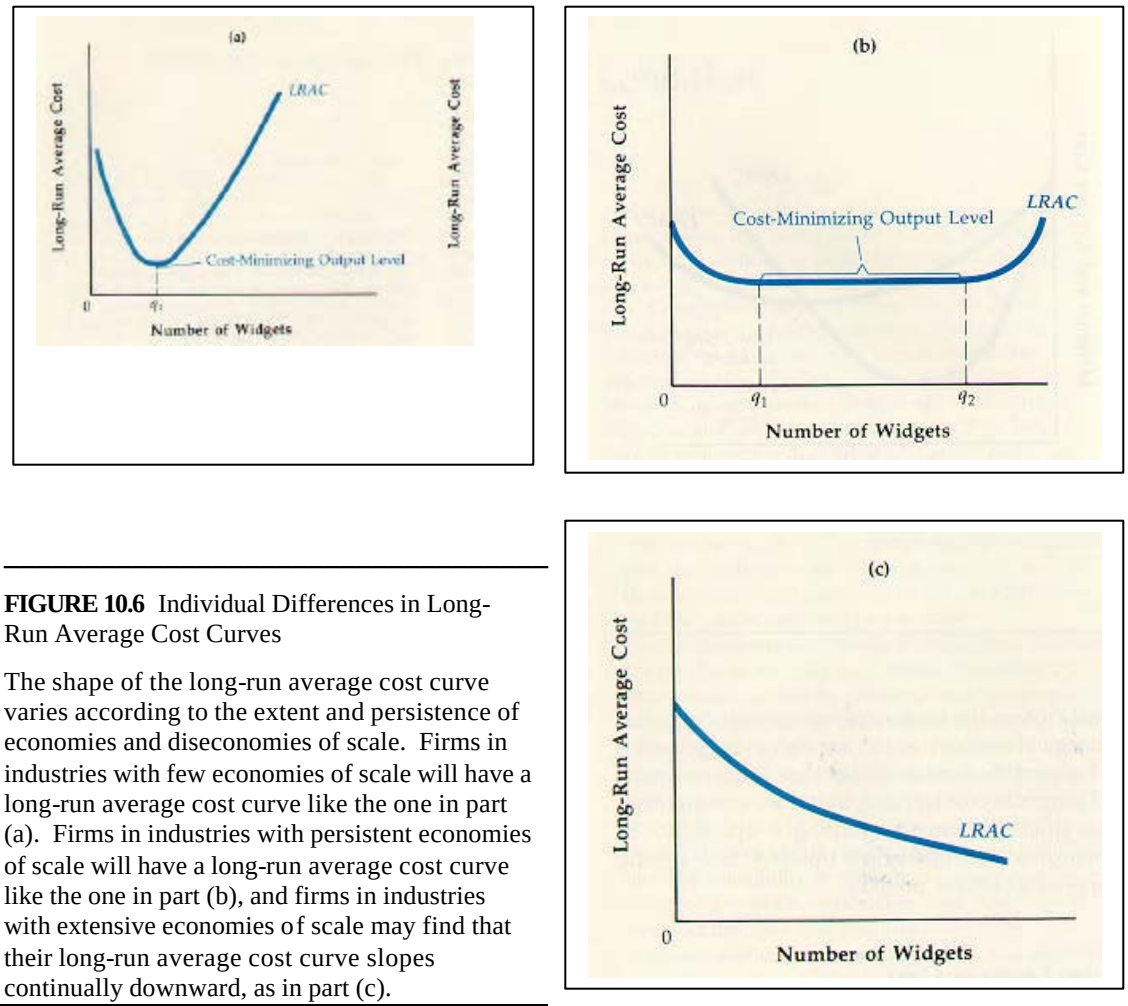


FIGURE 10.6 Individual Differences in Long-Run Average Cost Curves

The shape of the long-run average cost curve varies according to the extent and persistence of economies and diseconomies of scale. Firms in industries with few economies of scale will have a long-run average cost curve like the one in part (a). Firms in industries with persistent economies of scale will have a long-run average cost curve like the one in part (b), and firms in industries with extensive economies of scale may find that their long-run average cost curve slopes continually downward, as in part (c).

Shifts in the Average and Marginal Cost Curves

The average cost curves we have just described all assumed that the prices for resources remain constant. This is a critical assumption. If those prices change, so will the average cost curves. The marginal cost curve may shift as well, depending on the type of average cost—variable or fixed—that changes.

Thus if the price of a variable input—such as the wage rate of labor—rises, the firm's average total cost will rise along with its average variable cost (AFC + AVC = ATC), shifting the average total cost curve. The firm's marginal cost curve will shift as well, for the additional cost of producing an additional unit must rise with the higher labor cost (see Figure 10.7(a)). If a fixed cost like insurance premiums rises, average total cost will also rise, shifting the average total cost curve, as in Figure 10.7(b). The short-run marginal cost curve will not shift, however, because marginal cost is unaffected by fixed cost. The marginal cost curve is derived from variable costs only.

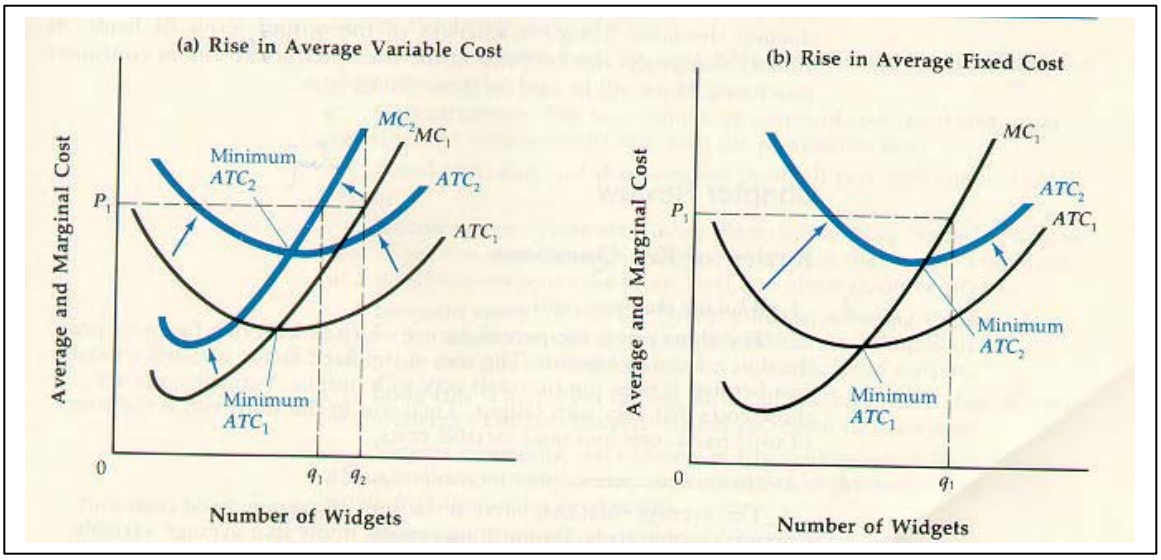


FIGURE 10.7 Shifts in Average and Marginal Costs Curves

An increase in a firm's variable cost (part (a)) will shift the firm's average total cost curve up, from ATC_1 to ATC_2. It will also shift the marginal cost curve, from MC_1 to MC_2. Production will fall because of the increase in marginal cost. By contrast, an increase in a firm's fixed cost (part (b)) will shift the average total cost curve upward from ATC_1 to ATC_2, but will not affect the marginal cost curve. (Marginal cost is unaffected by fixed cost.) Thus the firm's level of production will not change.

Because changes in variable cost affect a firm's marginal cost, they influence its production decisions. As we saw in an earlier chapter, a profit-maximizing firm selling at a constant price will produce up to the point where marginal cost equals price ($MC = P$). At a price of P_1 in Figure 10.7(a), then, the firm will produce q_2 widgets. After an increase in variable costs and an upward shift in the marginal cost curve, however, the

firm will cut back to q_1 widgets. At q_1 widgets price again equals marginal cost. The cutback in output has occurred because the marginal cost of producing $q_2 - q_1$ widgets now exceeds the price. In other words, an increase in variable cost results in a reduction in a firm's output.

Because a shift in average fixed cost leaves marginal cost unaffected, the firm's profit-maximizing output level remains at q_1 (see Figure 10.7(b)). The firm may make lower profits because of its higher fixed cost, but it cannot increase profits by either expanding or reducing output.

This analysis applies to the short run only. In the long run all costs are variable, and changes in the price of any resource will affect a firm's production decisions. Long-run changes in the output levels of firms, of course, change the market price of the final product as well as consumer purchases. More will be said on those points later.

MANAGER'S CORNER: **How Debt and Equity Affect Executive Incentives**

The cost structure that a firm faces is not *given* to the firm by some divine being. It emerges from the decisions made by managers, and their decisions depend critically upon the incentives they face, and managers' decisions depend on a number of factors. Here, we stress the importance of a firm's financial structure in shaping managers' incentives and their firms' cost structure.

The ideal firm is one with a single owner who produces a lot of stuff with no resources, including labor. Such a firm would be infinitely productive. It would totally avoid agency costs, or those costs that are associated with shirking of duties and the misuse, abuse, and overuse of firm resources for the personal benefit of the managers and workers who have control of firm resources. Agency costs can be expected to show up in lost output and a smaller bottom line for the firm. However, such an ideal firm cannot possibly exist.

The world we all do business in is one in which firms often need more funds for investment than one person can generate from his or her own savings or would want to commit to a single enterprise. Any single owner, if the business is even moderately successful, typically has to find ways of encouraging others to join the firm as owners or lenders (including bondholders, banks, and trade creditors).

Therein lies the source of many firms' problems, not the least of which is that a firm's expansion can give rise to the agency costs that a single-person firm would avoid. Managers and workers can use the expanding size of the firm as a screen for their shirking. The addition of equity owners (partners or stockholders) can dilute the incentive of any one owner to monitor what the agents do. Hence, as the firm expands, the agency costs of doing business can erode, if not totally negate, any economies of scale achieved through firm expansion.

One of the more important questions any single owner of a growing firm must face is, "How will the method of financing growth -- debt or equity -- affect the extent of

the agency cost?" Given that agency costs will always occur with expanding firms, how can the combination of debt and equity be varied to minimize the amount of costs from shirking and opportunism? That question is really one dimension of a more fundamental one, "How can the financial structure affect the firm's costs and competitiveness?"

In this short chapter, the eye of our focus is on debt, but that is only a matter of convenience of exposition, given that any discussion of debt must be juxtaposed with some discussion of equity as a matter of comparison, if nothing else. We could just as easily draw initial attention to equity as a means of financing growth. In fact, debt and equity are simply two alternative categories of finance (subject to much greater variation in form than we are able to consider here) available to owners. Owners need to search for an "optimum combination," given the features of both.

Debt and Equity as Alternative Investment Vehicles

By debt, of course, we mean funds, or the principal, that must be repaid fully at some agreed-upon point in the future and on which regular interest payments must be made in the interim. The interest rate is simply the annual interest payment divided by the principal. Also, we must note that in the event the firm gets into financial problems, the lenders have first claim on the firm's remaining assets.

By equity, or stock, we mean funds drawn from people who have ultimate control over the disposition of firm resources and who accept the status of residual claimants, which means a return on investment (which is subject to variation) will be paid only after all other claims on the firm have been satisfied. That is to say, the owners (stockholders) will not receive dividends until after all required interest payments have been met; the owners are guaranteed nothing in the form of repayment of their initial investments. Obviously, owners (stockholders) accept more risk on their investment than do lenders (or bondholders).[2]

Having outlined our intentions for this chapter, does it matter whether a firm finances its investments by debt or equity?[3] You bet it does (otherwise we must wonder why the two broad categories of finance would ever exist). The most important feature of debt is that the payments, both the payoff sum and the interest payments, are fixed. This is important for two reasons. One reason is the obvious one -- it enables firms to attract funds from people who want security and certainty in their investments. The modern aphorism, "different strokes for different folks," if followed in the structuring of financial

---

[2]We recognize that debt and equity come in a variety of forms. Common and preferred stock are the two major divisions of equity. Debt can take a form that has the "look and feel" of equity. For example, the much-maligned "junk bonds" often carry with them rights of control over firm decisions and may also be about as risky as common stock. In order to contain the length of this chapter, we consider only the two broad categories, and we will encourage readers to consult finance texts for more details on financial instruments. However, readers should recognize that variations in the type of debt and equity could help overcome some of the problems with each that are discussed in this chapter.

[3] For a more complete discussion of answers to this question, see Michael C. Jensen and William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," Journal of Financial Economics, vol. 3 (October 1976), pp. 305-360.

instruments, can mean lower costs of investment funds, growth, and competitiveness. Debt attracts funds from people who get their "strokes" from added security.

Fixed payments on debt are more important for our purposes for another reason: If the firm earns more than the required interest payments on any given investment project, the residual goes to the equity owners. If the company fails because of investments gone sour, then the firm is limited in its liability to lenders to the amount of their loans. If the firm is forced to liquidate its assets and the sale is insufficient to cover the debt, then it's simply going to be a sad day for the lenders (as well as stockholders, who will get nothing). The lenders can claim only what is left from the sale. That's it. Any profit remaining after all expenses have been covered doesn't have to be shared with the lenders. The remaining profits go to the equity stakeholders.

Clearly, the nature of debt biases, to a degree (depending on the exact features), the decision making of the owners, or their agent-managers, toward seeking risky investments, ones that will likely carry high rates of return. These high rates will, no doubt, incorporate a premium for risk taking, but they can also provide equity owners with an opportunity for a premium residual, given that they get what is left after the interest payments are deducted from high returns. If a firm borrows funds at a 10 percent interest rate, for example, and invests those funds in projects that have an expected rate of return of 12 percent, the residual left for the equity owners will be the difference, 2 percent. If, on the other hand, the funds are invested in a much riskier project that has a rate of return of 18 percent, then the residual that can be claimed by the equity owners is 8 percent, four times as great as the first case.

Granted, the project with the higher rate has a risk premium built into it (or else everyone investing in the 12 percent projects would direct their funds to the 18 percent projects, causing the rate of returns in the latter to fall and in the former to rise). However, notice that much of that additional risk is imposed on the lenders. They are the ones who must fear that the incurred risk will translate into failed investments (which is what risk implies). But they are not the ones who are compensated for the assumed risk they bear. Indeed, once a lender has made a loan, the managers can extend their indebtedness with more venturesome investments, increasing the risk imposed on the original lenders.

As a general rule, the greater the indebtedness, the greater incentive managers have to engage in risky investments. Again, this is because much of the risk is imposed on the lenders and the benefits, if they materialize, are garnered by the equity owners.

It should surprise no one that as a firm takes on more debt, lenders will become progressively more concerned that they will lose some or all of their investments. As a consequence, lenders will demand compensation in the form of higher interest payments, which reflect a risk premium. Those lenders who fear that the firm will continue to expand its indebtedness after they make the initial loans will also seek compensation prior to the rise in indebtedness by way of a higher interest rate. To keep interest costs under control, firm managers will want to find ways of making commitments as to how much indebtedness the firm will incur, and they must make the commitments believable, or else higher interest rates will be in the making. Again, we return to a reoccurring

theme in this book: managers' reputations for credibility have an economic value. In this case, the value emerges in lower interest payments.

Lenders, of course, will seek to protect themselves from risky managerial decisions in other ways. They may seek, as they often do, to obtain rights to monitor and even constrain the indebtedness of the firms to whom they make loans. Managers also have an interest in making such concessions because, although their freedom of action is restricted in one sense, they can be compensated for the accepted restrictions in the form of interest rates that are lower than otherwise. Firm managers are granted greater freedom of action in another respect; they are given a greater residual with which they can work (to add to their salary and perks, if they have the discretion to do so; extend the investments of the firm; or increase the dividends for stockholders).

Lenders may also specify the collateral the firm must commit. Lenders will not be interested in just any form of collateral. They will be most interested in having the firm pledge "general capital," or assets that are resaleable, which means that the lenders can potentially recover their invested funds. Lenders will not be interested in having "specific capital," or assets that are designed only for their given use inside a given firm. Such assets have little, if any, resale market.

Of course, firm assets are often more or less "general" or "specific," which means they can be better or worse forms of collateral. A firm can pledge assets with "specific capital" attributes. However, managers must understand that the more specific the asset (the narrower the resale market), the greater the risk premium that will be tacked onto the firm's interest rate, and the lower the potential residual for the equity owners.

Lenders will also have a preference for lending to those firms that have a stable future income stream and that can be easily monitored. The more stable the future income, the lower the risk of nonpayments of interest. The more easily the firm can be monitored, the less likely managers will be able to stick creditors with uncompensated risks. The more willing lenders are to lend to firms, the greater the likely indebtedness.

Electric utility companies have been good candidates for heavy indebtedness, because their markets are protected from entry by government controls and regulations, what they do is relatively easily measured, and their future income stream can be assumed to be relatively stable. Accordingly, their interest rates should be relatively low, which should encourage managers to take on additional debt just so that equity owners can claim the residual for themselves. (At this writing, the deregulation of electric power production is underway in a few states, which allows open entry into the generation of electricity. We should expect deregulation to lead to a higher risk premium in interest rates, although the price of electricity can be expected to fall for consumers with increased competition for power sales.)

Incentives in the S&L Industry

The incentives of indebtedness are dramatically illustrated in the biggest financial debacle of modern times, the dramatic rise in savings and loan bank failures of the 1980s. The S&L industry was established in the 1930s to ensure that the savings of individuals,

who effectively loaned their funds to the S&Ls, could be channeled to the housing industry (a concentrated focus of S&L investment portfolios that in itself added an element of risk, especially since housing starts vary radically with the business cycle). S&Ls were in a position to loan money for housing that was up to 97 percent from their depositors and only three percent from the owners (given reserve and equity requirements). Such a division, of course, made the S&L owners eager to go after high-risk but high-return projects. They could claim the residual from what was then a fixed interest payment on deposits.

When interests rates began to rise radically with the rising inflation rates of the late 1970s, alternative market-based forms of saving became available – not the least of which were money-market and mutual funds, which were unrestricted in the rates of return they could offer savers. As a consequence, savings started flowing out of S&Ls, which greatly increased the pressure on S&Ls to hike, when they were freed to do so, the interest rates on their deposits and to offset the higher interest rates by searching out investments that were risky but carried high rates of returns.

The S&Ls' incentive for risky investment was heightened by the fact that depositors' incentives to monitor the loans were severely muted by federal deposit insurance, which effectively assured the overwhelming majority of all depositors that they would lose nothing if all their S&L loans went sour.

To compensate for these perverse incentives, the federal government closely monitored and regulated the investments of the S&Ls through 1982. But that year, S&Ls were given greater freedom to pursue high-risk investments at the same time the protection to depositors was increased. The result was that which should have been predicted from the simple thought that if you give enough people a large enough temptation, many will succumb. S&Ls went after the high-risk/high-return -- and high residual -- investments. The S&Ls that made the risky investments were in a position to pay high interest rates, drawing funds from other more conservative S&Ls. In order to protect their deposit base, conservative S&Ls had to raise their interest rates, which meant that they, too, had to seek riskier investment, all of which led to a shock wave of risky investment spreading through the S&L/development industry.

Unfortunately, many of those investments did what should have been expected by their risky nature: they failed. The government had to absorb the losses and then return to doing that which it had done before 1982 -- closely monitor the industry and more severely restrict the riskiness of the investments (given that it was unwilling to give depositors greater incentives to monitor their S&Ls).

Clearly, fraud was a part of the S&L debacle. Crooks were attracted to the industry.[4] However, the debacle is a grand illustration of how debt can, and did, affect management decisions. It also enables us to draw out a financial/management principle: If owners want to control the riskiness of their firms' investments, they had better look to how much debt their firms accumulate. Debt can encourage risk taking, which can be

[4]See William K. Black, Kitty Calavita, and Henry N. Pontell, "The Savings and Loan Debacle of the 1980s: White-Collar Crime or Risky Business?" Law & Policy, vol. 17, no. 1 (Jan. 1995).

"good" or "bad," depending on whether the costs are considered and evaluated against the expected return.

Why then would the original equity owners ever be in favor of issuing more shares of stock and bringing in more equity owners with whom the original owners would have to share the residual? Sometimes, of course, the original owners are unable to provide the additional funds in order for the firm to pursue what are known (in an expectation sense) to be profitable investment projects. The original owners can figure that while their *share* of firm profits will go down, the *absolute level* of the residual they claim will go up. A 60 percent share of $100,000 in profits beats 100 percent of $50,000 in profits any day.

Another less obvious reason is that the additional equity investment can reduce the risk that the lenders face with loans to the firm. This means that the equity owners can claim a greater residual due to the fact that firm interest payments can fall with the reduction in the risk premium.

Often investment projects require a combination of specific and general capital to be used together. Consider, for example, the predicament of a remodeling firm that uses specially designed pieces of floor equipment (which may have little or no market value outside of the firm) as well as trucks that can easily be sold in well-established used truck markets. The investment projects can be divided according to the interests of the two types of investors. The equity owners can be called upon to take the risk associated with the floor equipment while the lenders are called upon to provide the funds for the trucks. Indeed, the lender might not even make the loan for the general part of the investment without equity owners taking the specific part precisely because the general investment would have limited value (or would carry undue risk) without the specific capital investment. (There may be no reason for the trucks if the firm has no floor equipment to work with.)

The original owners can also have an interest in selling a portion of their ownership share because, by doing so, they can reduce the overall risk of their full portfolio of investments by reinvesting the proceeds elsewhere, indeed, spreading their investments among a number of firms. If the original owners held their full investments in the firm, and refused to sell off a portion, then they might be "too cautious" in the choice of investments they would want the firm to pursue -- too reluctant to take the risky investments that can be the more rewarding endeavors.

By selling a portion of their interest in the firm, the original owners can actually change the direction of the firm's investment projects, and its growth, and can make the firm more profitable -- which translates into greater wealth for the original owners. The original owners can do this by lowering their (risk) costs by way of spreading their investments, and then by taking on more risky but more profitable investments in the original firm. Again, the financial structure of the firm is important -- and it can matter to management policies and to the bottom line.

Finance Professor Michael Jensen argues there is another reason for indebtedness for some firms: The interest payments on the debt can tie the hands -- or reduce the discretionary authority -- of managers who might otherwise engage in opportunism with

their firms' residual.[5] If a firm has little debt, then the managers can have a great deal of funds, or residual, to do with as they please. They can use the residual to provide themselves with higher salaries and more perks. They can also use the funds to contribute to local charities that may have little impact on their firm's business (they may have a warm heart for the cause they support or they may only want to take credit for being charitable with their firms' funds). They may also use the funds to expand (without the usual degree of scrutiny) the scope and scale of their firms, thereby giving reason for higher salaries and more perks (since size and executive compensation tend to go together) for themselves.

The investment projects the managers choose may indeed be profitable. The problem is that if the funds were distributed to the stockholders, the stockholders could find even more profitable investments (and even more worthy charitable causes).

As industries mature (or reach the limits of profitable expansion), the risk of managers "misusing" firm funds can grow. There may be few opportunities for managers to reinvest the earnings in their own industry. They may then be tempted to use the "excess residual" to fulfill some of their own personal flights of managerial fancy (give to charitable causes or pad their pockets), or reinvest the funds in other industries which may, or may not, have a solid connection to the original firm's core activities. Because of the additional costs of centralization and coordination of the investments across industries, the stock prices of mature companies can become depressed.

How can the firm be disgorged of the residual? Jensen suggests through indebtedness: the greater the indebtedness, the smaller the residual, and the less waste that can go up in the smoke of managerial opportunism. Jensen argues that one of the reasons for firm takeovers by way of "leveraged buyouts," which means heavy indebtedness, is that the firm is then forced to give up the residual through higher interest payments. Again, the hands of the agent-managers are tied; their ability to misuse firm funds is curbed. The value of the firm is enhanced by the indebtedness, mainly because it reduces the discretion of managers who have been misusing the funds. And managers can misuse their discretion in counterproductive ways, not the least of which is by diversifying the array of products and services provided on the grounds that diversity can smooth out the company's cash flows over the various cycles that go with the products and services. As Al Dunlap recognizes, "The flaw in that thinking is that shareholders are quite able to diversify on their own, thank you. Management doesn't have to do that for them."[6] But management does have to pass back the cash flow to the shareholders or, as the case may be, lenders.

---

[5] Michael C. Jensen, "Eclipse of the Public Corporation," Harvard Business Review (September-October 1989), pp. 64-65.

[6] Al Dunlap and Bob Andelman, Mean Business: How I Save Bad Companies and Make Good Companies Great (New York: Times Books, 1996), p. 81.

Firm Maturity and Indebtedness

This all leads us to an interesting proposition. We should expect firm indebtedness to increase with the maturity of its industry. Firms in a mature industry have more stable future income streams. They can be more easily monitored, given people's experience in working with the firms and knowing how such firms operate and are inclined to misappropriate funds when they do. Also, by taking on more debt, firms in mature industries can alert the market to their intentions to rid themselves of their residual, and not misuse managerial discretion, all of which can drive up the price of the firm's stock to a point that could not otherwise be reached.

Of course, if firms in mature industries don't take on relatively more debt and managers continue to misuse the funds by reinvesting the residual in the mature industry or other industries, then the firm can be ripe for a takeover. Some outside "raider" will see an opportunity to buy the stock, which should be selling at a depressed price, paying for the stock with debt. The increase in indebtedness can, by itself, raise the price of the stock, making the takeover a profitable venture. However, if the takeover target is, because of past management indiscretions in investment, a disparate collection of production units that do not fit well together, the profit potential for the raiders is even greater. The firm should be worth more in pieces than as a single firm. The raiders can buy the stock at a depressed price, take charge, and break the company apart, selling off the parts for more than the purchase price. In the process, the market value of the "core business" should be enhanced.

* * * * *

The moral of this "Manager's Corner" should now be self-evident: The financial structure of firms matters, and it matters a great deal. The structure can affect managerial actions and determine policies. The structure can also determine whether the firm will be the subject of a takeover. The one great antidote for a takeover should be obvious to managers, but it is not always (as evident by the fact that takeovers are not uncommon): Firms should be structured, both in terms of their financial and internal policies, in such a way that the stock price is maximized. In that case, potential raiders will have nothing to gain by taking the firm over. The jobs of the executives and their boards will be secure. Of course, one of the primary functions of a board of directors is to monitor the executives and the policies that are implemented with an eye toward maximizing stockholder value. As we will see, those executives and their board that do not maximize the price of their stocks do have something to fear from corporate raiders. They have definite reason, as we will see, to denigrate the social value of corporate raiders and to foil the takeover efforts of the raiders.

Concluding Comments

Short- and long-run costs are important topics in the study of economics. In order to understand how competitive and monopolistic markets operate, we must first understand the firm's cost structure. In following chapters, we will combine the average and marginal cost curves described here with the demand curves described in earlier chapters. Within that theoretical framework, we will be able to compare the relative efficiency of

competitive and monopolistic markets, and the role of profits in directing the production decisions of private firms.

Review Questions

1. Complete the cost schedule shown below and develop a graph that shows marginal, average fixed, average variable, and average total cost curves.

Output Level	Total Fixed Costs	Total Variable Costs	Total Cost	Marginal Cost	Average Fixed Cost	Average Variable Cost	Average Total Cost
1	$200	$ 60					
2	200	110					
3	200	150					
4	200	180					
5	200	200					
6	200	230					
7	200	280					
8	200	350					
9	200	440					
10	200	550					

2. Explain why the intersection of the average variable cost curve and the marginal cost curve is the point of minimum average variable cost.
3. Suppose no economies or diseconomies of scale exist in a given industry. What will the firm's long-run average and marginal cost curves look like? Would you expect firms of different sizes to be able to compete successfully in such an industry?
4. Why would you expect all firms would eventually encounter diseconomies of scale?
5. Suppose the government imposes a $100 tax on all businesses, regardless of how much they produce. How will the tax affect a firm's short-run cost curves? Its short-run production?
6. Suppose the government imposes a $1 tax on every unit of a good sold. How will the tax affect a firm's short-run cost curves? Its short-run output?
7. Suppose interest rates fall, how will managers' incentives be affected and how will the firm's cost structure be affected?

APPENDIX

Choosing the Most Efficient Resource Combination – Isoquant and Isocost Curves

The cost curves developed in this and previous chapters were based on the assumption that the producer had chosen the most technically efficient, cost-effective combination of resources possible at each output level. That is, resources were fully employed, were producing as much as possible, and were used in the lowest-cost combination. The short-run average total cost curve, for example, was as low as it could be, given the availability and prices of resources.

How does the firm find the most efficient combination of resources? Most products and output levels can be produced with various combinations of resources. A given quantity of blue jeans can be produced with a lot of labor and little capital (equipment) or a lot of capital and little labor. In Figure 10.A1, a firm can produce 100 pairs of jeans a day with five different combinations of labor and machines. Combination *a* requires seven workers and ten machines; combination b, five workers and fifteen machines. (To keep output constant, the use of labor must be reduced when the use of machines is increased. If the use of both were increased, output would rise.)

Curves like the one in Figure 10.A1 are called isoquants. An isoquant curve (from the Greek words for "same quantity") is a curve that shows the various technically efficient combinations of resources that can be use to produce a given level of output. Different output levels have different isoquants. The higher the output level, the higher the isoquant curve, as shown in Figure 10.A2. For example, an output level of 100 pairs of jeans can be produced with the resource combinations shown on curve $1Q_1$. An output level of 150 pairs of jeans requires larger resource combinations, shown on curve $1Q_2$.

To understand how the firm determines its most efficient resource combination, we must remember that it operates under conditions of diminishing marginal returns. The firm will always produce in the upward sloping range of its marginal cost curve; and marginal cost increases because marginal returns decline. Therefore, given a fixed quantity of one resource as more of another resource is used, the additional output marginal product, of that resource must diminish.

Then, as each additional worker is eliminated in Figure 10.A1, the number of machines added to keep output constant at 100 pairs of jeans must rise—and that is just what happens. Notice that as the firm moves down curve *abcde*, using fewer and fewer workers, the curve flattens out. At the same time that the marginal product of machines diminishes, the marginal product of the remaining workers rises.

Suppose, for instance, that the daily wage of labor is $100, and the daily rental for a sewing machine is $20. With a daily budget of $600, a firm can employ six workers and no machines or thirty machines and no workers. Or it can combine labor and machinery in various ways. It can employ four workers at a total expenditure of $400 and add ten machines at a total expenditure of $200. Curve IC_1 in Figure 10.A3 shows the various combinations of workers and machines the firm could choose. This kind of

curve is called an isocost curve. An **isocost** (meaning "same cost") **curve** is a curve that shows the various combinations of resources that can be employed at a given total expenditure (cost) level and given resource prices.

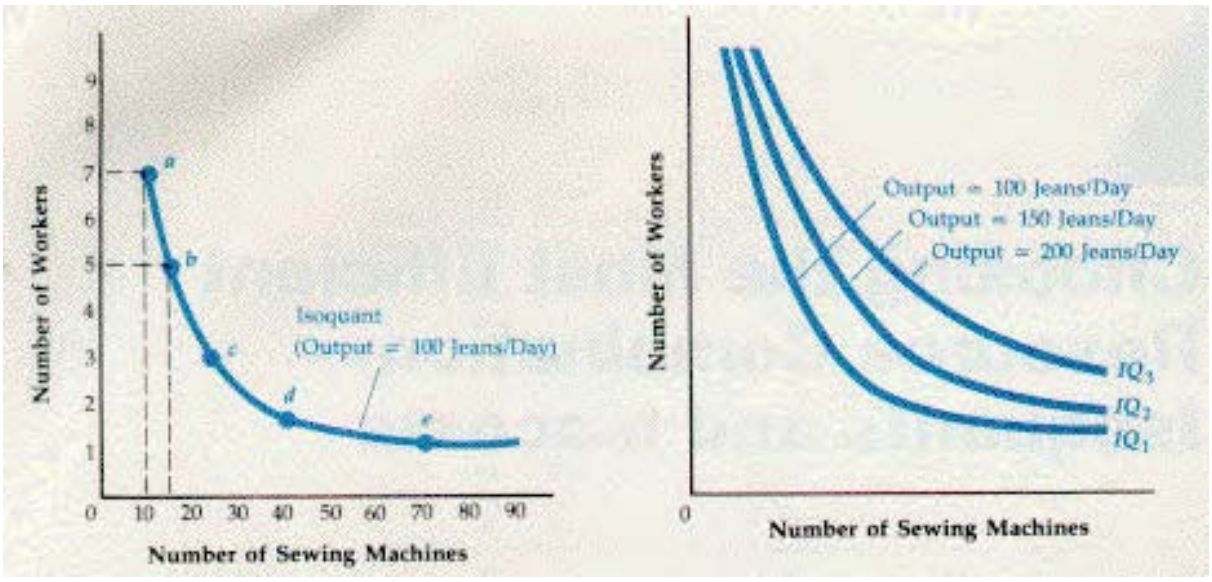


FIGURE 10.A1 Isoquant

A firm can produce one hundred pairs of jeans a day using any of the various combinations of labor and machinery shown on this curve. Because of diminishing marginal returns, more and more machines must be substituted for each worker who is dropped.

FIGURE 10.A2 Several Isoquants

Different output levels will have different insoquants. The higher the output level, the higher the isoquant.

We know, then, that the marginal product of resources differs with their level of use. To determine exactly which combination of resource should be employed to produce any given output level, however, we need to know not only the marginal product, but also the prices of labor and capital. The absolute prices of these resources will determine how much can be produced with any given expenditure. The relative prices will determine the most efficient combination.

There are different isocost curves for different output levels. The higher the output, the higher the isocost curve. As long as the prices of labor and capital stay the same, however, the various isocost curves for different output levels will be parallel to one another and will have the same downward slope.

Using both isoquant and isocost curves, we can determine the most efficient resource combination for a given expenditure level. Assuming a firm is on isocost curve IC_1 in Figure 10.A3 (which represents an expenditure of $600 per day), the most technically efficient and cost-effective combination of labor and capital will be point *a*, three workers and fifteen machines. At point *a* isocost curve IC_2 is tangent to isoquant

curve IQ_2. The firm is producing as much as it can -- 150 pairs of jeans a day -- with an expenditure of $600. If it produces the same amount but used more labor on more capital, it would move to a lower isoquant and a lower output level. A point *b* on curve IC_1, for instance, the firm would lower its production level from 150 to 100 pairs of jeans per day.

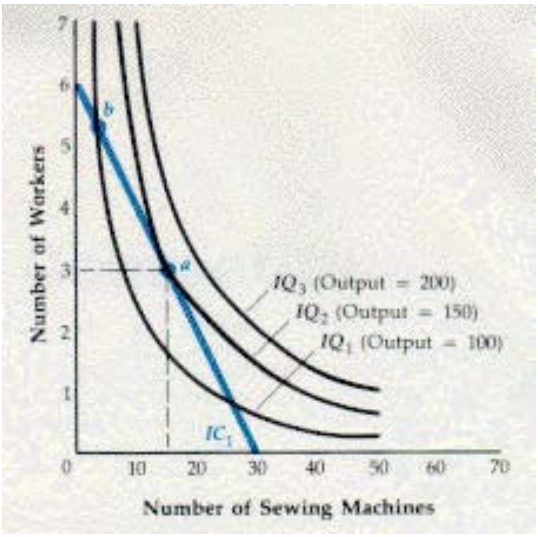


FIGURE 10.A3 Finding the Most Efficient combination of Resources

Assuming the dial wage of each worker is $100, and the daily rental on each sewing machine is $20, an expenditure of $600 per day will buy any combination of resources on isocost curve IC_1. The most cost-effective combination of labor and capital is point *a*, three workers and fifteen machines. At that point, the isocost curve is just tangent to isoquant IQ_2, meaning that the firm can product 150 pairs of jeans a day. If the firm chooses any other combination, it will move to a lower isoquant and a lower output level. At point *b* (on isoquant IS_1), it will be able to produce only 100 pairs of jeans a day.

Of course, with increased expenditures, the firm can move to a higher isocost curve. In figure 10.A4, as the firm's budget expands, its isocost curve shifts outward from IC_1 to IC_2 to IC_3. At the same time, the firm's most efficient combination of resources increases from *a* to *b* and then to *c*. As expenditures on resources rise, we can anticipate that beyond some point the increase in output will not keep pace with the increase in expenditure; at that point the marginal cost of a pair of jeans will rise.

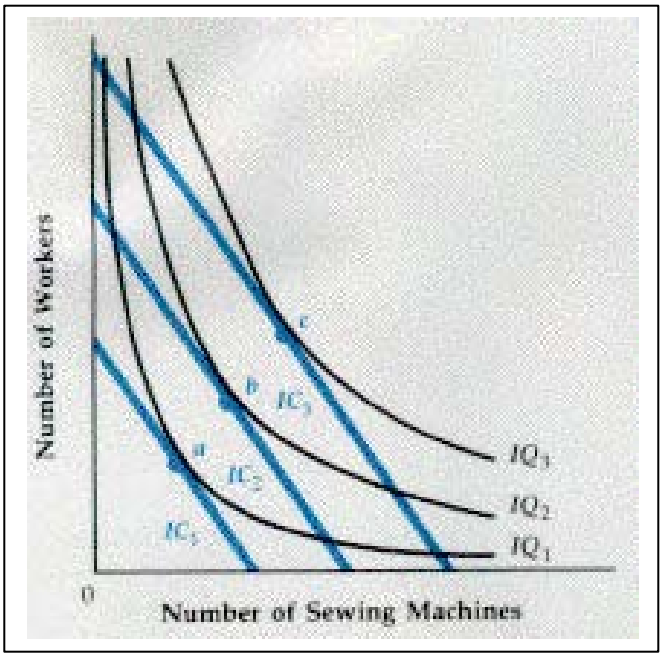


FIGURE 10.A4 The Effect of Increased Expenditures on Resources

An increase in the level of expenditures on resources shifts the isocost curve outward from IC_1 to IC_2. The firm's most efficient combination of resources shifts from point *a* to point *c*.

PERSPECTIVES: **Dealing with the Very Long Run**

Economic analysis tends to be restricted to either the short or the long run, for one major reason. For both periods, costs are known with reasonable precision. In the short run, firms know that beyond some point, increases in the use of a resource (for example, fertilizer) will bring diminishing marginal returns and rising marginal costs. They also know that with increased use of all resources, certain economies and diseconomies of scale can be expected over the long run. Given what is known about the technology of production and the availability of resources, economists can draw certain conclusions about a firm's behavior and the consequences of its actions.

As economists look further and further into the future, however, they can predict less about a firm's behavior and its consequences in the marketplace. Less is known about the technology and resources of the distant future. In the very long run, everything is subject to change—resources themselves, their availability, and the technology for using them. The **very long run** is the time period during which the technology of production and the availability of resources can change because if invention, innovation, and discovery of new technologies and resources.

By definition, the very long run is, to a significant degree, unpredictable. Firms cannot know today how to make use of unspecified future advances in technology. A hundred years ago firms had little idea how important lasers, satellites, airplanes, and computers would be to today's economy. Indeed, many products taken for granted today were invented or discovered quite by accident. Edison developed the phonograph while attempting to invent the light bulb. John Rock developed the birth control pill while studying penicillin, Charles Goodyear's development of vulcanization, and Wilhelm Roentgen's invention of the x-ray—all were accidents. All had economic consequences that could not have been predicted.

Not all inventions or innovations are accidental, and we can know something about the very long run. Firms have some idea of the value of investments in research and development. Research on substitute resources can yield improvements in productivity that translate into cost reductions. Research on new product designs will yield more attractive and useful products. There will be failures as well—research projects that accomplish little or nothing—but over time, the rewards of research and development can exceed the costs.

Because of the risks involved in research and development, some firms may be expected to fail. In the very long run, they will not be able to keep up with the competition in product design and productivity. The will not adjust sufficiently to changes in the market and will suffer losses. The computer industry provides many examples of firms that tried to build a better machine, but could not keep pace with the rapid technological advances of competitors.

Proponents of a planned economy see the uncertainty of the very long run as an argument for government direction of the nation's development. They stress that competitors often do not know what other firms are doing. Therefore they need guidance in the form of government subsidies and tax penalties to ensure that the nation's long-term goals are achieved.

Proponents of the market system agree that it is difficult to look ahead to the very long run, but they see the uncertainties as an argument for keeping production decisions in the hands of firms. Private firms have the economic incentive of profit to stay alert to changes in market conditions, and they can respond quickly to changes in technology and resources. Government control might slow the adjustment process.

CHAPTER 11

Firm Production under Idealized Competitive Conditions

Economists understand by the term market, not any particular market place in which things are bought and sold, but the whole of any region in which buyers and sellers are in such free intercourse with one another that the prices of the same goods tend to equality, easily and quickly.

Augustin Cournot

Preceding chapters dealt separately with the two sides of markets, consumers and producers. We devised graphic means of representing consumer preferences (the demand curve) and producer costs (average and marginal cost curves). This chapter brings demand and cost analysis together in order to examine the way in which individual firms react to consumer demand in competitive markets. Our focus will be on a highly competitive market structure called perfect competition. We will investigate an intriguing question: at the limit, how much can competitive markets contribute to consumer welfare?

We will not attempt to give a full description of a real-world competitive market setting. Because markets are so diverse, such a description would probably not be very useful. Our aim is rather to devise a theoretical framework that will enable us to *think* about how markets work in general, as a constructive behavioral force. Although our model cannot tell much that is specific about real-world markets, it will provide a basis for predicting the general direction of changes in market prices and output. Through its analysis, we should gain a deeper understanding of the meaning of market efficiency.

Perfect competition is only one of four basic market structures. The other three, and the detrimental effects of their restrictions on competition, are the subjects of following chapters.

The Four Market Structures

Markets can be divided into four basic categories, based on the degree of competition that prevails within them -- that is, on how strenuously participants attempt to outdo, and avoid being outdone by, their rivals. The most competitive of the four market structures is perfect competition.

Perfect Competition

As we stressed much earlier in the book, perfect competition represents an ideal degree of competition. Perfect competition can be recognized by the following characteristics:

1. There are many producers in the market, no one of which is large enough to affect the going market price for the product. All producers are price takers, as opposed to price searchers or price makers (see the Perspective on the subject below).
2. All producers sell a homogeneous product, meaning that the goods of one producer are indistinguishable from those of all others. Consumers are fully knowledgeable about the prices charged by different producers and are totally indifferent as to which producer they buy from.
3. Producers enjoy complete freedom of entry into and exit from the market—that is, entry and exit costs are minimal, although not completely absent.
4. There are many consumers in the market, no one of whom is powerful enough to affect the market price of the product. Like producers, consumers are price takers.

As we have seen before, the demand curve facing the individual perfect competitor is not the same as the demand curve faced by all producers. The market demand curve slopes downward, as shown in Figure 11.1(a). The demand curve facing an individual producer -- price taker -- is horizontal, as in Figure 11.1(b). This horizontal demand curve is perfectly elastic. That is, the individual firm cannot raise its price even slightly above the going market price without losing all its customers to the numerous other producers in the market or to other producers waiting for an opportunity to enter the market. On the other hand, the individual firm can sell all it wishes at the going market price. Hence it has no reason to offer its output at a lower price. The markets for wheat and for integrated computer circuits, or computer chips, are both good examples of real-world markets that come close to perfect competition.

Pure Monopoly

Pure monopoly: A single seller of a product for which there are no close substitutes. Protected from competition by barriers to entry into the market. The barriers to entry into the monopolist's market will be described in the next chapter. For now, we will simply note that because the monopolistic firm does not have to worry about competitors undercutting its price, it can raise its price without fear that customers will move to other producers of the same product or similar products. All the pure monopolist has to worry about is losing customers to producers of distantly related products.

Since the monopolist is the only producer of a particular good, the downward-sloping market demand curve [Figure 11.1(a)] is its individual demand curve. Unlike the perfect competitor, the monopolistic firm can raise its price and sell less, or lower its price and sell more. The critical task of the pure monopolist is to determine the one price-quantity combination of all price-quantity combinations on its demand curve that maximizes its profits. In this sense the pure monopolist is a price searcher. The best (but not perfect) real-world examples of a pure monopoly are regulated electric-power companies, which dominate in given geographical areas, and the government's first-class postal system.

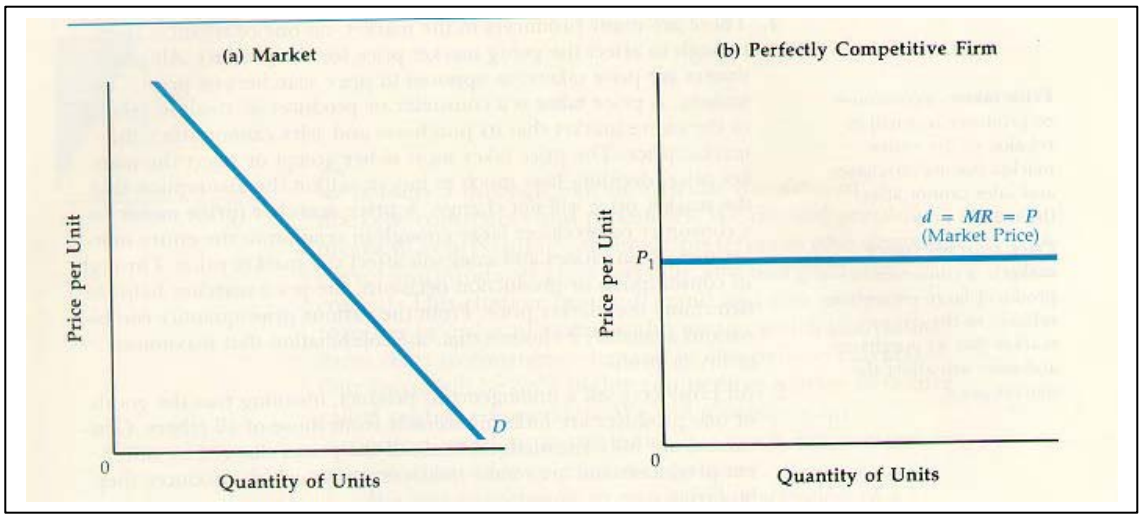


FIGURE 11.1 Demand Curve Faced by Perfect Competitors

The market demand for a product part (a) is always downward sloping. The perfect competitor is on a horizontal, or perfectly elastic, demand curve [part (b)]. It cannot raise its price above the market price even slightly without losing its customers to other producers.

Monopolistic Competition

Monopolistic competition is a market composed of a number of producers whose products are differentiated and who face highly elastic, but not perfectly elastic, demand curves.

A monopolistically competitive market can be recognized by the following characteristics:

1. There are a number of competitors, producing slightly different products.
2. Advertising and other forms of nonprice competition are prevalent.
3. Entry into the market is not barred but is restricted by modest entry costs, mainly overhead.
4. Because of the existence of close substitutes, customers can turn to other producers if a monopolistically competitive firm raises its price. Because of brand loyalty, the monopolistic competitor's demand curve still slopes downward; but it is fairly elastic [see Figure 11.2).

The market for textbooks is a good example of monopolistic competition. Most subjects are covered by two or three dozen textbooks, differing from one another in content, style of presentation, and design.

PERSPECTIVE: **Price Takers and Price Searchers**

Perfect competition is an extreme degree of competition, so much so that many students are understandably concerned about its relevance. They often ask, "If there are few market structures that even closely approximate perfect competition, why bother to study it?" The question is a good one and not altogether easy to answer. There are few markets that come close to having numerous producers of an identical product with complete freedom of entry and exit. Markets for agricultural commodities and for stocks and bonds are probably the closet markets we have to perfect competition, but still the products are not always *completely* identical, and entry and exit costs abound in most markets. Even wheat sold by a Kansas wheat farmer us not always viewed the same as wheat sold by a Texas wheat farmer.

How can sense be made of perfect competition? The answer is remarkably simple. We know that under the conditions of competition specified, certain results follow. We can logically (with the use of graphs and mathematics) derive them, and the results are developed in this and the following chapter. One conclusion drawn is that in perfect competition each firm will extend production until the marginal cost of producing the last unit equals the price paid by the consumer. That conclusion *necessarily* follows. As we will see, it is mathematically valid. The strict (extreme) assumptions about the nature of perfect competition assure that.

The demanding conditions for perfect competition are rarely met. We nevertheless cannot conclude that under less demanding competitive conditions, competitive results would not be observed. [see the *Perspectives* on contestable markets on page 240.) For example, it may be that the number of producers is not "numerous" that the products sold by all producers are not completely "identical," and that there are costs to moving in and out of markets. Nonetheless, individual producers may act *as if* the conditions of perfect competition are met. Individual producers may still act as if they have no control over market price or that there are so many other actual or potential producers that it is best to think in terms of the other producers being numerous"—in which case many of the predicted results of perfect competition may be still observed in the less-than-perfect markets.

For these reasons, many economists often talk not about *perfect competitors* but about price takers (who may or may not fit exactly the description of perfect competitors). *Price takers* are sellers who do not believe they can control the market price by varying their own production levels. They simply observe the market price and either accept it (and produce accordingly, to the point where marginal cost and marginal revenue and price are equal) or reject it (and go into some other business). The price taker is someone who acts *as if* his or her demand curve is horizontal (perfectly elastic, more or less). He or she is therefore someone who assumes the marginal revenue on each unit sold is constant (and equal to the price)—and that the marginal revenue curve is horizontal and the same as the firm's demand curve.

The *price searcher* stands in contrast to the price taker. *Price searchers* are sellers who have some control over the market price. Price searchers have monopoly power due to the fact that they can alter production and thereby market supply sufficiently to change the price. The individual price searcher's task is not simply to accept or reject the current market price, but (like the monopolist) to "search" through the various price-quantity combinations on his or her downward sloping demand curve with the intent upon maximizing profits. As we will see in the following chapter, the marginal revenue and demand curves of the price searcher are no longer the same. (Exactly where the monopolist's marginal revenue curve lies in relation to the demand curve will be discussed in detail in the next chapter).

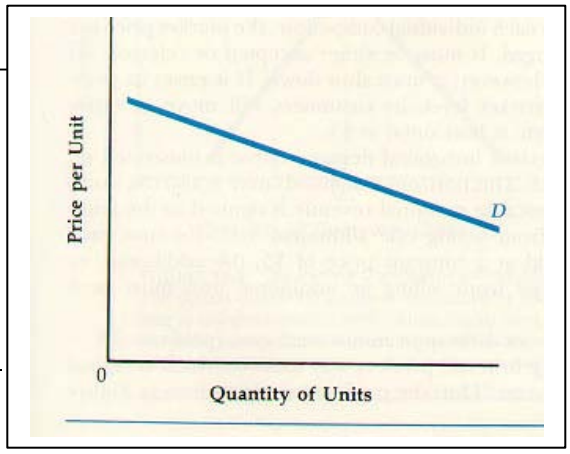


FIGURE 11.2 Demand Curve Faced by a Monopolistic Competitor

Because the product sold by the monopolistically competitive firm is slightly different from the products sold by competing producers, the firm faces a highly elastic, but not perfectly elastic, demand curve.

Oligopoly

An **oligopoly** is a market composed of only a handful of dominant producers—as few as two—whose pricing decisions are interdependent. Oligopolists may produce either an identical product (like steel) or highly differentiated products (like automobiles). Generally the barriers to entry into the market are considerable, but the critical characteristic of oligopolistic firms is that their pricing decisions are interdependent. That is, the pricing decisions of any one firm can substantially affect the sales of the others. Therefore, each firm must monitor and respond to the pricing and production decisions of the other firms in the industry. The importance of this characteristic will become clear in a following chapter.

Table 11.1 summaries the characteristics of the four market structures.

The Perfect Competitor's Production Decision

As we learned earlier, the market price in a perfectly competitive market is determined by the intersection of the supply and demand curves. If the price is above the equilibrium price level, a surplus will develop forcing competitors to lower their prices. If the price is below equilibrium, a shortage will emerge, pushing the price upward [see Figure 11.3(a)]. Given a market price over which it has no control, how much will the individual perfect competitor produce?

The Production Rule: MC = MR

Suppose the price in the perfectly competitive market for computer chips \$5 (P_1 in Figure 11.3). For each individual competitor, the market price is given, that is, cannot be changed. It must be either accepted or rejected. If the firm rejects the price, however, it must shut down. If it raises its price even slightly above the market level, its customers will move to other competitors.) Demand, then, is horizontal at \$5.

Table 11.1 Characteristics of the Four Market Structures.

	Number of Firms	Freedom of Entry	Type of Product	Example
Perfect competition	Many	Very easy	Homogeneous	Wheat, Computers, and Gold
Pure monopoly	One	Barred	Single product	Public utilities and Postal service
Monopolistic competition	Many	Relatively easy	Differentiated	Pens, Books, Paper, and Clothing
Oligopoly	Few	Difficult	Either standardized or differentiated	Steel, Light bulbs, Cereal, and Autos

The firm's perfectly elastic horizontal demand curve is illustrated on the right side of Figure 11.3. This horizontal demand curve is also the firm's marginal revenue curve, because marginal revenue is defined as the additional revenue acquired from selling one additional unit. Because each computer chip can be sold at a constant price of $5, the additional, or marginal, revenue acquired from selling an additional unit must be constant at $5.

Because profit equals total revenue minus total cost (profit = $TR = TC$), the profit-maximizing firm will produce any unit for which marginal revenue exceeds marginal cost. Thus the profit-maximizing firm in Figure 11.3(b) will produce and sell q_1 units, the quantity at which marginal revenue equals marginal cost ($MR = MC$). Up to q_1, marginal revenue is greater than marginal cost. Beyond q_1, all additional computer chips are unprofitable: the additional cost of producing them is greater than the additional revenue acquired [with the small "*q*" being used to remind you that the output individual producer in Figure 11.3(b) is a small fraction of the output for the market, designated by a capital "*Q*" in Figure 11.3 (a)].

Changes in Market Price

The perfectly competitive firm produces where MC = MR, both of which are equal to price. Thus the amount the firm produces depends on market price. As long as market demand remains constant, the individual firm's demand, and its price, will also remain constant. If market demand and price increase, however, the individual firm's demand and price will also increase.

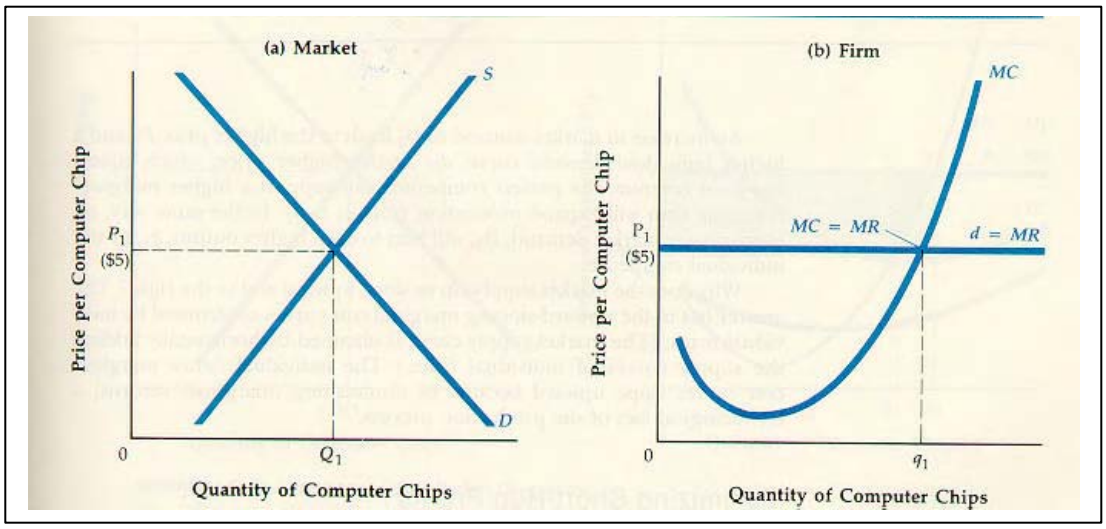


FIGURE 11.3 The Perfect Competitor's Production Decision

The perfect competitor's price is determined by market supply and demand [part (a)]. As long as marginal revenue (MR), which equals market price, exceeds marginal cost (MC), the perfect competitor will expand production [part (b)]. The profit-maximizing production level is the point at which marginal cost equals marginal revenue (price).

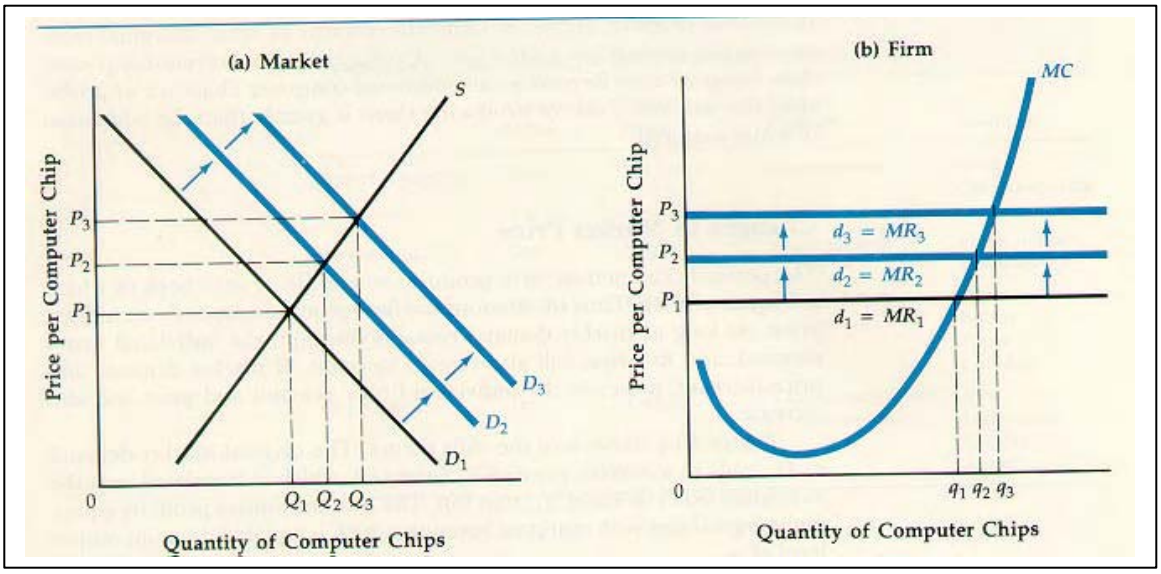


FIGURE 11.4 Change in the Perfect Competitor's Market Price

If the market demand rises from D_1 to D_3 [part (a)], the price will rise with it, from P_1 to P_3. As a result, the perfectly competitive firm's demand curve will rise, from d1 to d3 [part (b)].

Figure 11.4 (above) shows how the shift occurs. The original market demand of D_1 leads to a market price of P_1 [part (a)], which is translated into the individual firm's demand, d_1 [part (b)]. The firm maximizes profit by equating marginal cost with marginal revenue, which is equal to d_1, at an output level of q_1.[1]

An increase in market demand to D_2 leads to the higher price P_2 and a higher individual demand curve, d_2. At this higher price, which equals marginal revenue, the perfect competitor can support a higher marginal cost. The firm will expand production from q_1 to q_2. In the same way, an even greater market demand, D_3, will lead to even higher output, q_3, by the individual competitor.

Why does the market supply curve slope upward and to the right? The answer lies in the upward-sloping marginal cost curves confronted by individual firms. (The market supply curve is obtained by horizontally adding the supply curves of individual firms.) The individual firm's marginal cost curves slope upward because of diminishing (marginal) returns, a technological fact of the production process.

Maximizing Short-Run Profits

Can perfect competitors make an economic profit? The answer is yes, at least in the short run. To see this point, we must incorporate the average and marginal cost curves developed in the last chapter into our graph of the perfect competitor's demand curve, as in Figure 11.5(b). [Figure 11.5(a) shows the market supply and demand curves.) As before, the producer maximizes profits by equating marginal cost with price, rather than by looking at average cost. That is exactly what the perfect competitor does. The firm produces q_2 computer chips because that is the point at which marginal revenue curve (which equals the firm's demand curve crosses the marginal cost curve. At that intersection, marginal revenue of the last unit sold equals its marginal cost. If less were produced that q_1, the marginal cost would be less than the marginal revenue, and profits would be lost. Similarly, by producing anything more than q_2, the firm incurs more additional costs (as indicated by the marginal

---

[1] To prove this statement, first we note that

$$TR = \bar{P}Q$$

Then we define short-run total cost to be a function of output:

$$SRTC = C(Q)$$

Next, we define profits π to be

$$\pi = TR - SRTC = \bar{P}Q - C(Q)$$

Differentiating with respect to Q and equating with 0, we then obtain

$$\frac{d\pi}{dQ} = \bar{P} - \frac{dC(Q)}{dQ} = 0$$

$$\bar{P} = \frac{dC(Q)}{dQ}$$

Since $\frac{dC(Q)}{dQ}$ = SRMC, profits are maximized when SRMC $= \bar{P}$.

cost curve) than it receives in additional revenue (as indicated by the demand curve, which beyond q_2 is below the MC curve).

At q_2 (and anywhere else), the firm's profit equals total revenue minus total cost ($TR - TC$). To find total revenue, we multiply the price, P_1 (which also equals average revenue) by the quantity produced, q_2 ($TR = P_{1q2}$). Graphically, total revenue is equal to the area of the rectangle bounded by the price and quantity, or $0P_{1aq2}$.

Similarly, total cost can be found by multiplying the average total cost of production (ATC) by the quantity produced. The ATC curve shows us that the average total cost of producing q2 computer chips is ATC_1. Therefore total cost is ATC_{1q2}, or the rectangular area bounded by $0ATC_1bq_2$. The profits of the company are therefore $P_{1q2} - ATC_1q_2$, which is the same, mathematically, as $q_2(P_1 - ATC_1)$. This quantity corresponds to the area representing total revenue, OP_1aq_2, minus the area representing total cost, $0ATC_1bq_2$. Profit is the shaded rectangle bounded by ATC_1P_1ab.

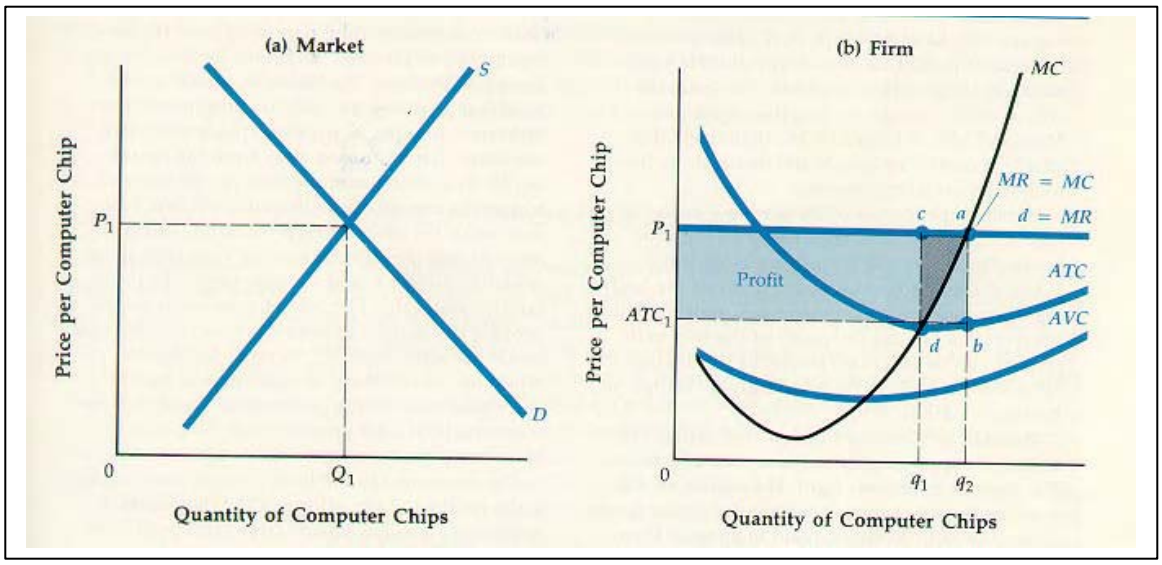


FIGURE 11.5 The Profit-Maximizing Perfect Competitor

The perfect competitor's demand curve is established by the market-clearing price [part (a)]. The profit-maximizing perfect competitor will extend production up to the point where marginal cost equals marginal revenue (price), or point *a* in part (b). At that output level—*q2*—the firm will earn a short-run economic profit equal to the shaded area ATC_1P_{1ab}. If the perfect competitor were to minimize average total cost, it would produce only q_1, losing profits equal to the darker shaded area *dca* in the process.

The perfect competitor does not seek to produce the quantity that results in the lowest average total cost. That quantity, q_1, is defined by the intersection of the marginal cost curve and the average total cost curve. If it produced only q_1, the firm would lose out on some of its profits, shown by the darker shaded area *dca*. (Suppose the firm is producing at q_1. If it expands production to q_2, it will generate P_1 times $q_2 - q_1$ in extra revenue (price times the additional units sold), an amount represented graphically by the area q_1daq_2.)

Naturally, profit-maximizing firms will attempt to minimize their costs of production. That does not mean they will produce at the point of minimum average total cost. Instead, the will try to employ the most efficient technology available and to minimize their payments for resources. That is they will attempt to keep their cost curves as low as possible. But given those curves, the firm will produce where *MC = MR*, not where *ATC* is at its lowest level. Managers who cannot distinguish between those two objectives will probably operate their businesses on a less profitable basis than they might—and will risk being run out of business.

Minimizing Short-Run Losses

In the foregoing analysis the market-determined price was higher than the firm's average total cost, allowing it to make a profit. Perfect competitors are not guaranteed profits, however. The market price may not be high enough for the firm to make a profit. Suppose, for example, that the market price is P_1, below the firm's average total cost curve [see Figure 11.6). Should the firm still produce where marginal cost equals marginal revenue (price)? The answer, for the short run, is yes. As long as the firm can cover its variable cost, it should produce q_1 computer chips.

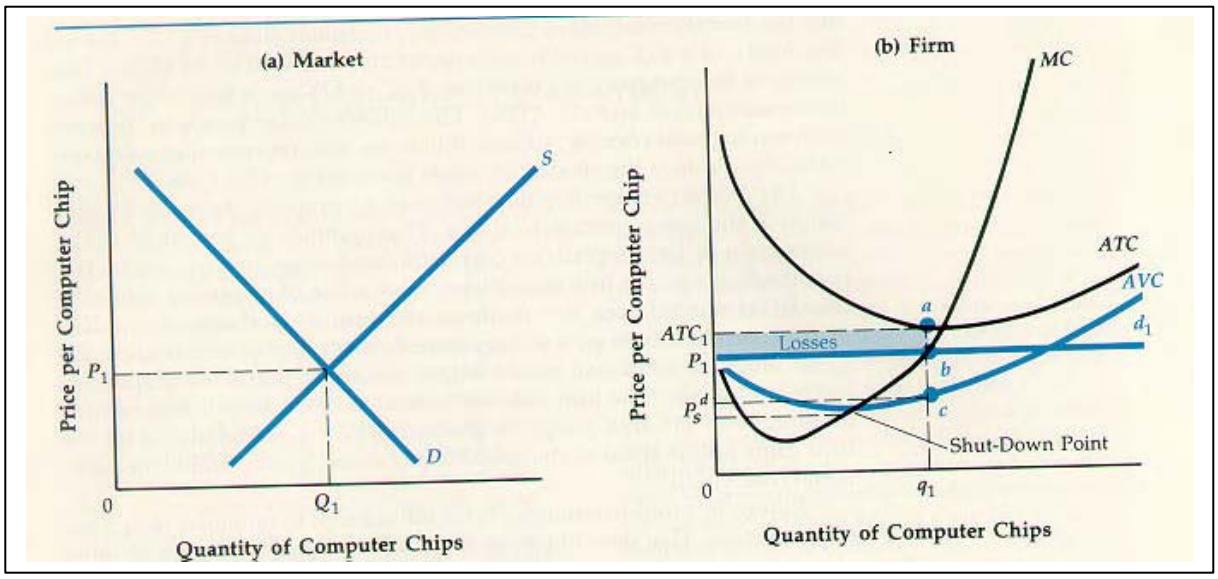


FIGURE 11.6 The Loss-Minimizing Perfect Competitor

The market-clearing price [part (a)] establishes the perfect competitor's demand curve [part (b)]. Because the price is below the average total cost curve, this firm is losing money. As long as the price is above the low point of the average variable cost curve, however, the firm should minimize its short-run losses by continuing to produce where marginal cost equals marginal revenue [price or point *b* in part (b)]. This perfect competitor should produce q_1 units, incurring losses equal to the shaded area P_1ATC_1ab. (The alternative would be to shut down, in which case the firm would lose all its fixed costs.)

It is true that the firm will lose money. Its total revenues are only P_{1q1}, or the area bounded y $0P_1bq_1$, whereas its total costs are ATC_1q_1, or the area $0ATC_1aq_1$, whereas its total

costs are ATC_1q_1, or the area $0ATC_1aq_1$. On the graph its total losses equal the difference between those two rectangular areas, the shaded area bounded by P_1ATC_1ab. Whether the firm incurs losses is not the relevant question, however. The real issue is whether the firm loses more money by shutting down or by operating and producing q_1 chips.

In the short run, the firm will continue to incur fixed costs even if it shuts down. If it is not earning any revenues, its losses will equal its total fixed costs. In the last chapter we saw that the average fixed cost of production is the vertical distance between the average variable cost and average total cost.

In short, as long as the price is higher than average variable cost—if the price more than covers the cost associated directly with production—the firm minimizes its short-run losses by producing where marginal cost equals marginal revenue. Only if the price dips below the low point of the average variable cost curve—where the marginal and average variable cost curves intersect—will the firm add to its losses by operating. The firm will shut down when price is at or below that point, P_s in Figure 11.6. At prices above that point, the firm simply follows its marginal cost curve to determine its production level. Above the average variable cost curve, then, the marginal cost curve is in effect the firm's supply curve. Therefore, if a perfect competitor produces at all, it produces in a range of increasing marginal cost—and diminishing marginal returns.

Our analysis has shown why, in the short run, fixed costs should be ignored. The relevant question is whether a given productive activity will add more to the firm's revenues than to its costs. Understanding this principle, businesses may undertake activities that superficially appear to be quite unprofitable. Some grocery stores stay open all night, even though the owners known they will attract few customers. If all costs, including fixed costs, are considered, the decision to operate in the early morning hours may seem misguided. The only relevant question facing the store manager is whether the additional sales generated are greater than the additional cost of light, goods sold, and labor. Similarly, many businesses that are obviously failing continue to operate, for by staying open they can at least cover a portion of their fixed costs—such as rent—that would still be due if they shut down. They stay open until their leases expire or until they can sell out.

Producing Over the Long Run

In the long run businesses have an opportunity to change their total fixed costs. If the market price remains too low to permit profitable operation, a firm can eliminate its fixed costs, sell its plant and equipment, or terminate its contracts for insurance and office space. If the market price is above average total cost, new firms can enter the market, and existing firms can expand their scale of operation. Such long-run adjustments in turn affect market supply, which affects price and short-run production decisions.

The Long-Run Effect of Short-Run Profits and Losses

When profits encourage new firms to enter an industry and existing firms to expand, the result is an increase in market supply, a decrease in market price, and a decrease in the profitability of individual firms. For example, in Figure 11.7(a), the existence of economic profits in the computer chip market means that investors can earn more in that industry than in some others. Some investors will move their resources to the computer chip industry. Because the number of producers increases, the supply curve shifts outward, expanding total production from Q_1 to Q_2 and depressing the market price from P_2 to P_1.

The expansion of industry supply and the resulting reduction in market price make the computer chip business less profitable for individual firms. The lower market price is reflected in a downward shift of the firm's horizontal demand curve, from d_1 to d_2 [see Figure 11.7(b)]. The individual firm reduces it output from q_2 to q_1, the intersection of the new marginal revenue (price/demand) curve with the marginal cost curve. Note that q_1 is also the low point of the average total cost curve. Here price equals average total cost, meaning that economic profit is zero. The firm is making just enough to cover its opportunity and risk costs, but no more.

Losses have the opposite effect on long-run industry supply. In the long run, firms that are losing money will move out of the industry, because their resources can be employed more profitably elsewhere. When firms drop out of the industry, supply contracts and total

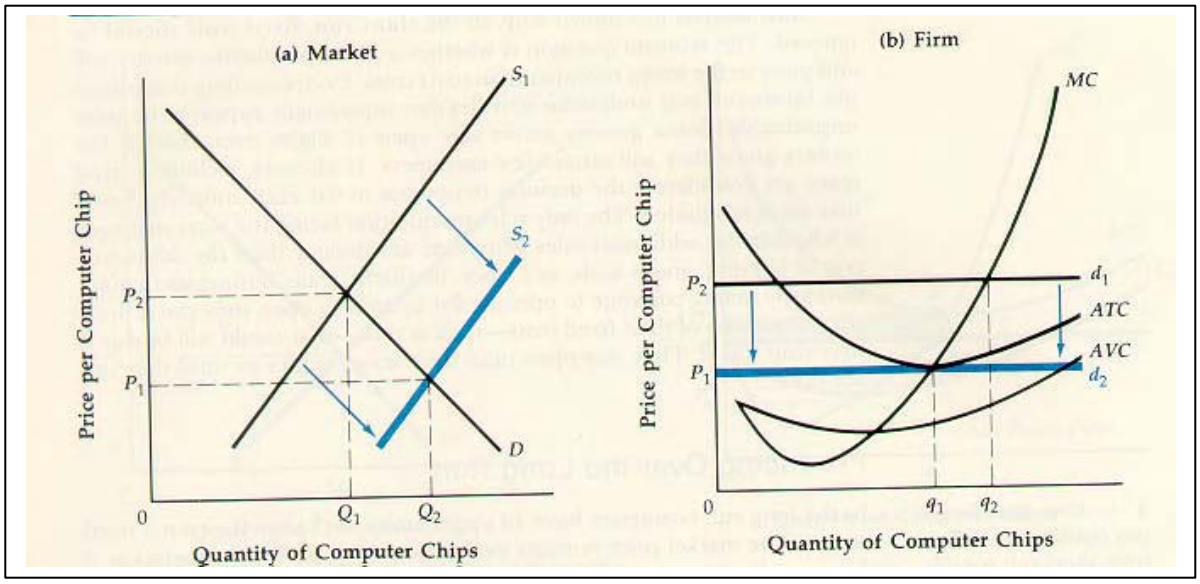


FIGURE 11.7 The Long-Run Effects of Short-Run Profits

If perfect competitors are making short-run profits, other producers will enter the market, increasing the market supply from S_1 to S_2 and lowering the market price, from P_2 to P_1 part (a). The individual firm's demand curve, which is determined by market price will shift down, from d_1 to d_2 [part (b)]. The firm will reduce its output from q_2 to q_1, the new intersection of marginal revenue (price) and marginal cost. Long-run equilibrium will be achieved when the price falls to the low point of the firm's average total cost curve, eliminating economic profit [price P_1 in (b)].

production falls, from Q_2 to Q_1 in Figure 11.8(a). As a result, the price of the product rises, permitting some firms to break even and stay in the business. Long-run equilibrium occurs when the price reaches P_2, where the individual firm's demand curve is tangent to the low point of the average total cost curve [Figure 11.8(b)]. The output of each remaining individual firm expands (from q_1 to q_2) to take up the slack left by the firms that have withdrawn. Again price and average total cost are equal, and economic profit is zero.

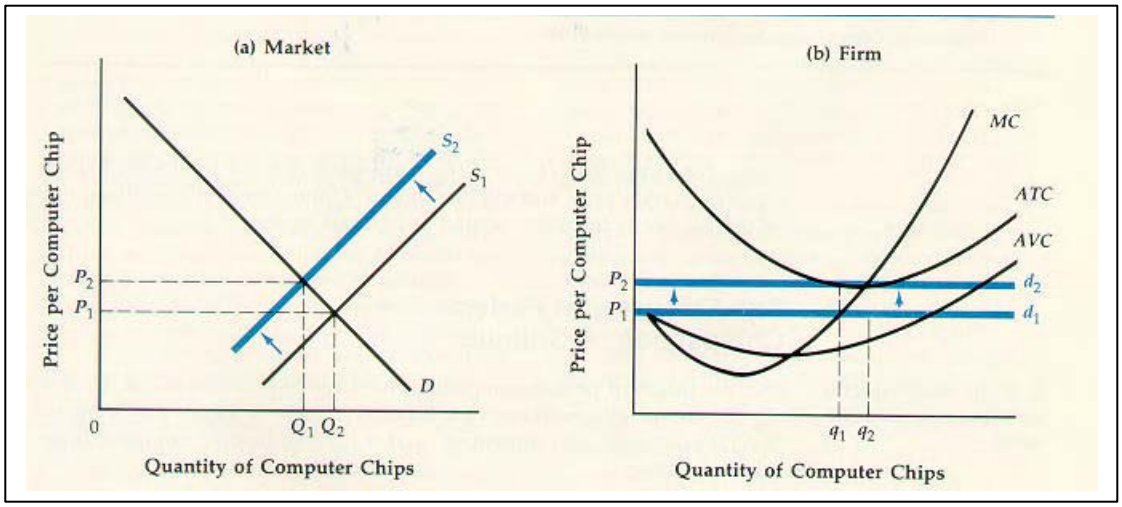


FIGURE 11.8 The Long-Run Effects of Short-Run Losses

If perfect competitors are suffering short-run losses, some firms will leave the industry causing the market supply to shift back from S_1 to S_2 and the price to rise, from P_1 to P_2 part (a). The individual firm's demand curve will shift up with price, from d_1 to d_2 [part (b)]. The firm will expand from q_1 to q_2, and equilibrium will be reached when price equals the low point of average total cost P_2, eliminating the firm's short-run losses.

The Effect of Economies of Scale

In the long run, competition forces firms to take advantage of economies of scale, if they exist. If expanding the use of resources reduces costs, the perfect competitor must expand. Otherwise, other firms will expand their scale of operation, increasing market supply and forcing the market price down. Any firm that does not expand its scale will be caught with a cost structure that is higher than the market price. In addition to mere self-preservation, the firm also has a profit incentive for expansion. If it expands before other firms, its lower average total cost will allow it to make greater profits for a short period of time.

Consider Figure 11.9, for instance. Initially the market is in short-run equilibrium at a price of P_2 [part (a)]. The individual firm is on cost scale ATC_1, producing q_1 chips and breaking even [part (b)]. If the firm expands its scale of operation and produces where its demand curve d_1 intersects the long-run marginal cost curve, it will make a profit equal graphically to the shaded area ATC_1P_2ab. That is the firm's incentive for expansion.

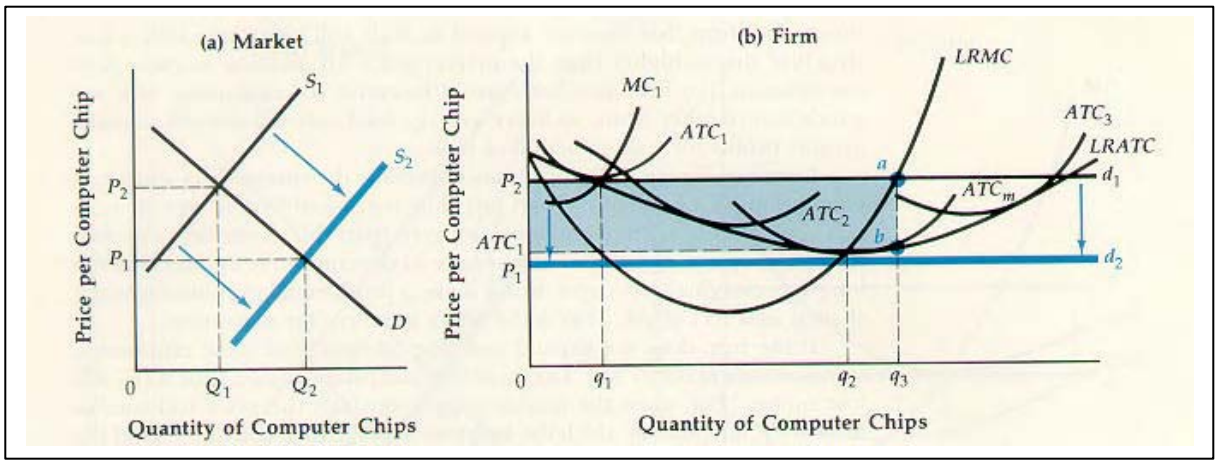


FIGURE 11.9 The Long-Run Effects of Economies of Scale

If the market is in equilibrium at price P_1 in part (a). and the individual firm is producing q_1 units on short-run average total cost curve ATC_1 [part (b)], firms will be just breaking even. Because of the profit potential represented by the shaded area ATC_1P_2ab, firms can be expected to expand production to q_3, where the long-run marginal cost curve intersects the demand curve (d_1). As they expand production to take advantage of economies of scale, however, supply will expand from S_1 to S_2 in part (a), pushing the market price down toward P_1, the low point of the long-run average total cost curve (*LRAC*). Economic profit will fall to zero. Because of rising diseconomies of scale, firms will not expand further.

If the firm does not expand and take advantage of these economies, some other firm surely will. Then any firm still producing on scale ATC_1 will lose money. For when the market supply expands the price will tumble toward P_1, the point at which the long-run average total cost curve (and the short-run curve ATC_m) are at a minimum, and both industry and firm profits are zero. Because of rising diseconomies of scale, firms will not be able to expand further. Any firm that tries to produce on a smaller or larger scale—for example, ATC_2 or ATC_3 -- will occur average total costs higher than the market price and will lose money. Ultimately it will be driven out of the market or forced to expand or contract its scale.

The Efficiency of Perfect Competition: A Critique

Our discussion of perfect competition has been highly theoretical. In real life, the competitive market system is not as efficient as the analysis may suggest. Several aspects of the competitive market deserve further comment from this perspective.

The Tendency Toward Equilibrium

Market forces are stabilizing: they tend to push the market toward one central point of equilibrium. To that extent the market is predictable, and to that extent it contributes to economic and social stability. In the real world price does not always move as smoothly toward equilibrium as it appears to do in supply and demand models. The smooth, direct

move to equilibrium may happen in markets where all participants, both buyers and sellers, know exactly what everyone else is doing. Often, however, market participants have only imperfect knowledge of what others are going to do, for one function of the market is to generate the pricing and output information people need to interact with one another.

In a world of imperfect information, then, prices may not and probably will not move directly toward equilibrium. Those who compete in the market will continually grope for the "best" price, from their own individual perspectives. At times sellers will produce too little and reap unusually high profits.

This process of groping toward equilibrium can be represented graphically by a supply and demand "cobweb" [see Figure 11.10). Most producers must plan their production at least several months ahead on the basis of prices received today or during the past production period. Farmers, for instance, may plant for summer harvest on the basis of the previous summer's prices. Suppose farmers got price P_1 for a bushel of wheat last year. Their planning supply curve, S, will encourage them to work for a harvest of only Q_1 bushels this year. Given that limited output and the rather high demand at price P_1, however, the price farmers actually receive is P_4. The price of P_4 in turn induces farmers to plan for a much larger production level, Q_3, the following year. The market will not clear for Q_3 bushels, however, until the price falls to P_2. The next year farmers plan for a price of P_2 and reduce their production to Q_2—which causes the price to rise to P_3. As you can see from the graph, instead of moving in a straight line, the market moves toward the intersection of supply and demand in a web-like pattern.

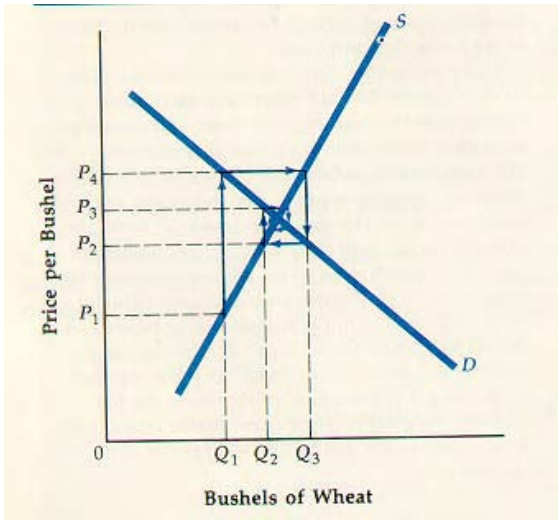


FIGURE 11.10 Supply and Demand Cobweb

Markets do not always move smoothly toward equilibrium. If current production decisions are based on past prices, price may adjust to supply in the cobweb pattern shown here. Having received price P_1 in the past, farmers will plan to supply only Q_1 bushels of wheat. That amount will not meet market demand, so the price will rise to P_4—inducing farmers to plan for a harvest of Q_3 bushels. At price P_4, however, Q_3 bushels will not clear the market. The price will fall to P_2, encouraging farmers to cut production back to Q_2. Only after several tries many farmers find the equilibrium price-quantity combination.

Surpluses and Shortages

Some critics complain that the market system creates wasteful surpluses and shortages. Although all resources are limited in quantity, a true market shortage can exist only if the going price is below equilibrium. Thus shortages can be eliminated by a price increase.

How much of an increase, theory alone cannot say. We do know, however, that market forces, if allowed free play, will work to boost the price and eliminate the shortage. That means, if course, that people of limited financial resources will be eliminated from the market—an enduring concern that motivates many government efforts to legislate market conditions.

Similarly, all surpluses exist because the going price is above equilibrium. Competition will reduce the price, eliminating the surplus. In the process, of course, some firms will be driver out of the market and into other, more productive activities. Others will be unable to keep their employees working full-time. A frequent criticism of the market system is that when this happens, workers have difficulty finding employment in other lines of production. Part of the problem, however, is that labor contracts, community custom, or minimum wage laws prevent wages from adjusting downward. If government controls prices—that is, if prices are not permitted to respond to market conditions-—surpluses and shortages will persist.

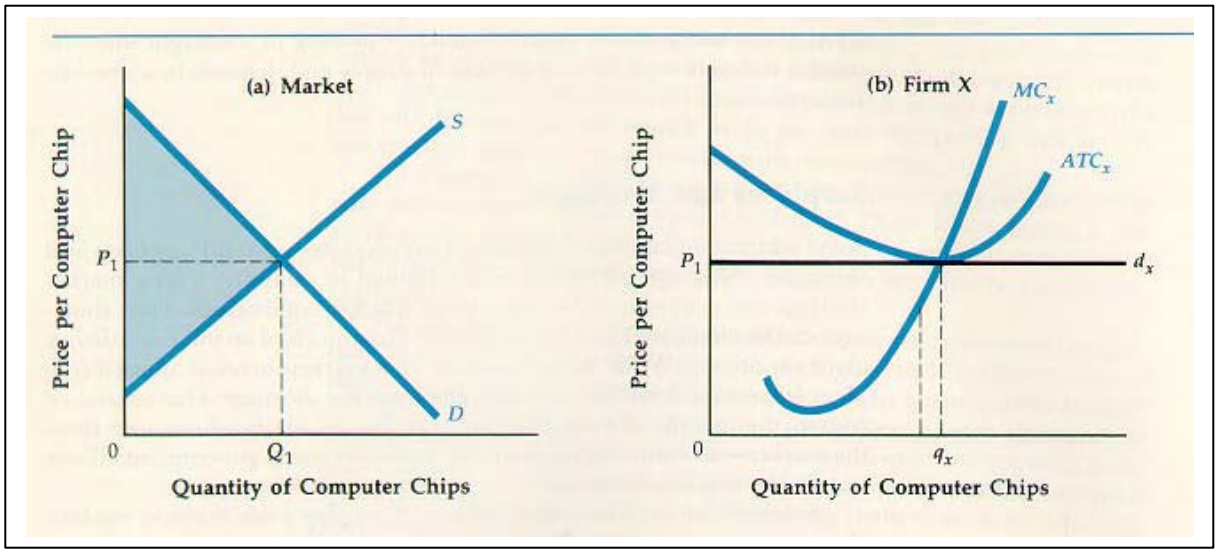
Marginal Benefit Versus Marginal Cost

Time lags, surpluses, and shortages notwithstanding, the competitive market can produce efficient results in one important sense. That is that the marginal benefit of the last unit produced equals its marginal cost ($MB = MC$). In Figure 11.11(a), for every computer chip up to Q_1, consumers are willing to pay a price (as indicated by the demand curve, D) greater than its marginal cost (as indicated by the industry supply curve, S). The difference between the price consumers are willing to pay—an objective indication of the product's marginal benefits—and the marginal cost of production is a kind of surplus, or net gain received from the production of each unit. The net gain is composed of two surpluses, consumer surplus and producer surplus. **Consumer surplus** is the difference between the total willingness of consumers to pay for a good and the total amount actually spent. In Figure 11.11(a) consumer surplus is the triangular area below the demand curve and above the dotted price line, P_1. **Producer surplus** is the difference between the minimum total revenue necessary to induce producers to supply Q_1 units of output and the actual total revenue received from selling that output. In Figure 11.11(a), producer surplus is the triangular area above the supply curve and below the dotted price line, P_1. By producing Q_1 units, the industry exploits all potential gains from production, shown graphically by the shaded triangular area in the figure. That net gain is brought about by the price that is charged, P_1—a price that induces individual firms to produce where the marginal cost of production equals the price, which is also equal to consumers' marginal benefit.

The marginal cost of production for each individual firm is also P_1, a fact that results in the production of Q_1 units at the minimum total cost. Parts (b) and (c) show the cost curves of two firms, X and Y. In competitive equilibrium, firm X produces q_x, units. Suppose that the market output were distributed between the firms differently. Suppose, for example, that firm X produced one computer chip less than q_x. To maintain a constant market output of Q_1, firm Y (or some other firm) would then have to expand production by one unit. The additional chip would force firm Y up its marginal cost curve. To Y, the marginal cost of the additional chip is greater than P_1, greater than X's marginal cost to produce it. Competition

forces firms to produce at a cost-effective output level and therefore minimizes the cost of producing at any given level of output.

Perfectly competitive markets are attractive for another reason. In the long run, competition forces each firm to produce at the low point of its average total cost curve. Firms must either produce at that point, achieving whatever economies of scale are available, or get out of the market, leaving production to some other firm that will minimize average total cost.



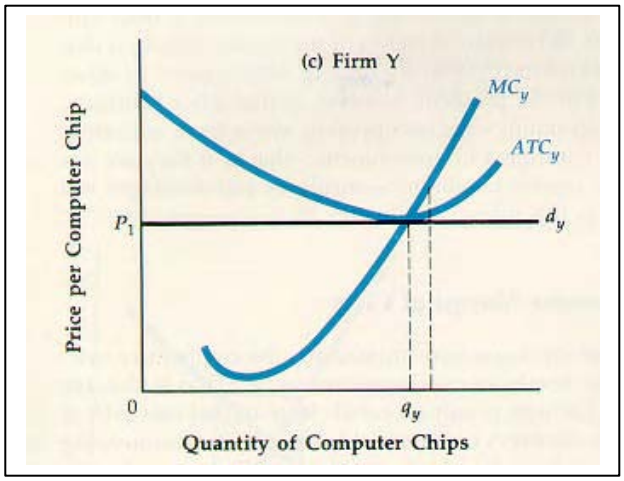


FIGURE 11.11 The Efficiency of the Competitive Market

Perfectly competitive markets are efficient in the sense that they equate marginal benefit [shown by the demand curve in part (a)] with marginal cost (shown by the supply curve). At the market output level, Q_1, the marginal benefit of the last unit produced equals the marginal cost of production. The gains generated by the production of Q_1 units—that is , the difference between cost and benefits—are shown by the shaded are in part (a).

The perfectly competitive market is also efficient in the sense that the marginal cost of production, P_1, is the same for all firms [parts (b) and (c)]. If firm X were to produce fewer than its efficient number of units, q_x, firm Y would have to produce more than its efficient number, q_y, to meet market demand. Firm Y would be pushed up its marginal cost curve, to the point where the cost of the last unit exceeds its benefits. But competition forces the two firms to produce to exactly the point where marginal cost equals marginal benefit, thus minimizing the cost of production.

Critics stress, however, that supply is based only on the costs firms bear privately. External costs like air, noise, and water pollution are not counted as part of the cost of production. If the external costs of pollution were counted, the firm's supply curve would be lower, S_2 instead of S_1 in Figure 11.12. If producers and consumers had to pay all the costs of production, only Q_1 units would be bought. In this sense, competition leads to overproduction of $Q_2 - Q_1$ units. The cost of producing these $Q_2 - Q_1$ chips is the area under the supply curve between Q_1 and Q_2, Q_1abQ_2. The benefit to consumers is the area under the demand curve, or the area Q_1acQ_2. The extent to which the cost of overproduction exceeds the benefits to consumers is shown by the shaded triangular area *abc*.

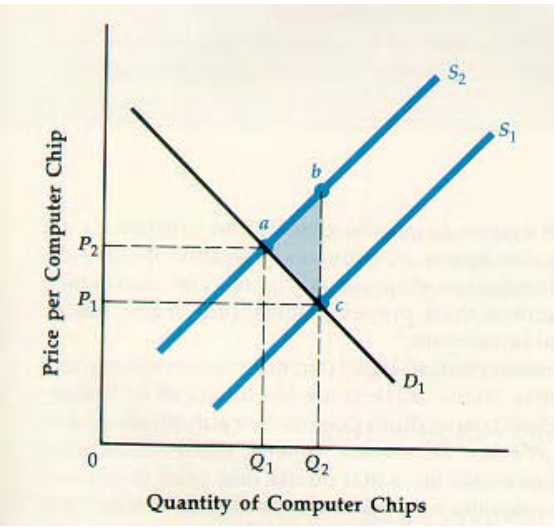


FIGURE 11.12 Inefficiency Caused by External Costs

If external costs equal to the vertical distance *bc* are not counted as costs of production, supply will be artificially high at S_1, and firms will overproduce by $Q2 - Q_1$ units. The inefficiency, or welfare loss, form such overproduction is shown by the shaded area *abc*, the amount by which the total cost of producing $Q_2 - Q_1$ units (shown by curve S_2) exceeds their total benefits (shown by the demand curve).

Critics of the market system stress also that its cost efficiencies are achieved within a specific distribution of resources of wealth, one that depends on the existing distribution of property rights. The distribution of economic power inherent in these property rights, they argue, has no particular ethical or moral significance.

Finally, critics of the market system argue that most real-world markets are not perfectly competitive. Actual markets are not inhabited by numerous firms producing standard commodities that can be easily duplicated by anyone who would like to enter the market. Indeed, many markets are inhabited by a few large, powerful firms that do not take price as a given. Many firms either are monopolies or possess a high degree of monopoly power. Demanders and suppliers are rarely as well informed as the model suggests. The model of perfect competition was never meant to represent all or even most markets. It is merely one of several means economists use to think about markets and the consequences of changes in market conditions and government policy.

Critics of the market system stress also that its cost efficiencies are achieved within a specific distribution of resources or wealth, one that depends on the existing distribution of property rights. The distribution of economic power inherent in these property rights, they argue, has no particular ethical or moral significance.

Contestable Markets

One of the most important developments in the study of markets is the theory of contestable markets.[2] The contestable market model stresses the importance of potential rather than actual competitors in a market. A market is deemed to be contestable if entry and exit are relatively easy. A market is perfectly contestable if entry is absolutely free and exit is costless. Free entry has a particular meaning in the theory of contestable markets; it means that new firms entering an industry are not at any cost disadvantage compared to existing firms in the industry. In other words, latecomers suffer no cost handicaps. Costless exit means that firms can leave the industry at any time and can recoup all costs incurred by entry.

A contestable market, then, is marked by ease of entry and exit and in that respect is similar to a perfectly competitive market. Like a perfectly competitive market, a contestable market will be characterized by zero economic profits in the long run. For a contestable market, however, we do not need a large number of firms and a homogeneous product. Indeed, multiproduct firms are possible in contestable markets. A contestable market may have only two or three firms operating in it. Moreover, those firms produce at rates of output where price is equal to marginal cost.

What brings about this result? Why do firms in contestable markets not produce and price at the monopoly equilibrium? The reason is entry and exit. If price is not equal to marginal cost, profit opportunities exist and new firms will quickly enter the market, causing existing firms to make losses. The potential competitors force the existing firms to produce where price equals marginal cost. A firm in a contestable market is always open to hit and run attacks from its potential competitors. They will therefore be forced to produce and sell at an output where price equals marginal cost and economic profits are zero. Any attempt to exploit market power will bring about entry into the market and the dissipation of all profits. The firms in the contestable market will be forced to operate as it they were in perfectly competitive markets.

A contestable market is depicted in Figure 11.14. Note that although only three firms are in the industry, they all produce where price equals marginal and average cost. For the industry as a whole, price is equal to the minimum on the long-run average total cost curve. Each firm produces one-third (q) of total industry output ($3q$). Production at an efficient rate of output and marginal cost pricing, then, do not require the atomistic markets of the perfectly competitive model. A perfectly contestable market will do.

What industries might this model fit? The air travel industry is one candidate. Many major markets are served by only two or three airlines. Yet if an airline with a dominant position in a particular regional market attempted to set price well above costs, entry would quickly follow. Airplanes can be shifted from one market or use to another with ease. New

---

[2] The basic model of a contestable market is presented in William J. Baumol, "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, 72 (March 1982), 1-15. For a critical analysis of the model, see William G. Shepherd, "'Contestability' vs. Competition," *American Economic Review* 74 (September 1984), pp. 572-587.

entrants do not appear to be at a cost disadvantage relative to existing firms. If the conditions for a contestable market were indeed met, then we would expect the air travel industry to be characterized by marginal cost pricing and zero economic profits. It is always difficult to determine whether or not price is equal to marginal cost; one indication that contestability characterizes the air travel industry is that prices do not appear to be higher in markets with fewer actual competitors. The zero-profit outcome also describes the air travel industry reasonable well.

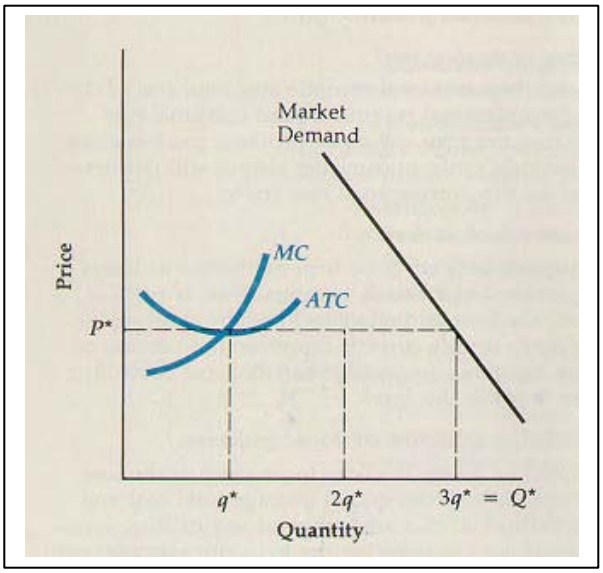


FIGURE 11.14 A Contestable Market

The market is composed of three firms, each producing output q^*, which minimizes average costs. Total industry output is $Q^* = 3q^*$. Any attempt by the three firms to reduce output and increase market price will lead to entry by new firms and the dissipation of profits.

MANAGER'S CORNER: **When Workers Would Want Their Bosses to Cut Their Pay**

In trying to manage a firm's production and cost properly, managers want to cater to many (not all) of their workers' wishes. What is more obvious than the desire of workers for higher salaries and wages? Certainly no sane person would deny that all workers would rather be paid more money rather than less, everything else equal. But everything else is seldom equal. For example, while workers may rather take home bigger paychecks with the work being held equal, they do not necessarily want a higher wage if it requires less pleasant or more difficult responsibilities.[3] But even for the same work, workers may prefer to be paid less money. Indeed, workers are better off because employers are constantly looking for, and succeeding in finding, ways to pay them less. This is a point you very likely haven't seen covered in your human resource studies.

[3]As explained in an earlier chapter, despite what they may say, most young and inexperienced MBA graduates would not want a job paying $200,000 immediately upon graduation. Such an employee would have to contribute at least $200,000 to firm revenues, which he or she, without experience, is not likely to be able to do. The expected value of a job with a much lower salary is likely to be higher, given the much higher probability of the new graduate keeping it.

In the analysis that follows, always keep in mind a key point in our above examples: workers can be better off even when they experience a cut in their monetary pay. Workers are better off with the lower pay than they were before because the only way the employer could reduce their pay (without reducing their ability to hire competent workers) is by substituting a fringe benefit that is worth even more than the lost pay. Even though monetary pay has been cut, total compensation has been increased. We advance the argument by showing that workers benefit more from a fringe the larger the reduction in their wage.

Our demonstration in this section is, admittedly, subject to one qualification. That qualification is that all workers have much the same preferences for fringe benefits. Even if workers in general are made better off when a fringe benefit is substituted for monetary compensation, it is possible that the fringe benefit is one that some workers do not value at all, or do not value as much as they do the loss of money income. This is a problem, however, that both workers and employers have a strong motivation to overcome. Workers will be more attracted to firms that offer the combination of fringe benefits and money wages that best conforms to their preferences. For example, we have noted many young workers seeking part-time employment to help pay for college will want most of their compensation in money wages with little in the way of fringe benefits. They need cash and most face low (or no) tax rates. In general, older workers in higher income tax brackets and with greater demand for medical care will want more of their compensation in the form of untaxed fringe benefits such as health insurance. Therefore we can expect employers who hire a lot of young part-time workers to offer fewer fringe benefits than employers who hire mostly adult full-time workers. Also, employers will find it to their advantage in competing for workers to offer a menu of fringe benefits from which workers can choose. The closer an employer can adjust the fringe benefit package to the preferences of the workers, the more the employer can save by paying lower money wages.[4]

But even if we assume that all workers benefit equally from the fringe benefits provided, can we really show that workers receive the greatest benefit when their wages are cut the most? What is to keep the employer from receiving all the advantage from fringe benefits that are worth more than they cost? Sure, an employer will provide fringe benefits that cost only $50 per worker if they are worth $100 to each worker. But if the employer then reduces each worker's money income by the full $100, which she could presumably do without losing any workers, where is the gain to workers? The answer is that if some way is found to save on the cost of hiring workers, competition will force employers to share some – but not all -- of those savings with workers.

For example, if one firm discovers a fringe benefit that lowers the cost of workers, other firms will find advantage in providing that benefit also. With workers becoming less costly to firms in general, they will be more aggressively sought out, and the competition between firms will prevent any one firm from lowering the wages of workers by the full value of the new fringe benefit. Also, even if the fringe benefit could only be provided by the one firm, so workers did not become more valuable to other firms, competition would

[4] Over half of big American companies with 100 or more workers give those workers a chance to tailor their benefits. This means that a worker covered by a spouse's health insurance can opt out of health insurance and have the savings applied to, say, dental insurance or car insurance. The purpose of giving workers the option is, naturally, cost control (The Economist, December 21, 1996, p. 91).

still prevent the one firm providing the benefit from cutting the money wage by the full value of the benefit. Assume that the firm did initially attempt to capture all the value of a fringe benefit by lowering the wage by its full value. The result would be that workers now cost the firm a lot less than before, and this cost advantage would make it profitable to hire more workers. But the only way to hire more workers is to bid them away from other activities, which means bidding at least some away from other firms. This can only be done by increasing the wage back up -- not to the point where it was before the fringe benefit was added, but high enough so that the total compensation is higher with the fringe benefit than it was before, even though the money wage is lower than before.

You are excused, however, if you are not yet convinced that when a fringe benefit is provided the gain to workers is greater the larger the reduction in their money wage. Our verbal discussion of the effect of fringe benefits is not sufficiently precise to convincingly establish the connections between those benefits, the money wage, and the gain to workers. The best way to get at these connections is by returning to the demand and supply curves used earlier. With those curves we have already seen that if the fringe benefit is worth providing (its value is greater than its cost), then the wages of workers will fall by more than the cost of the fringe benefit (the employer gains) but by less than its value to workers (the workers gain). Illustrating this important point again will set the stage for understanding why the bigger the wage cut the better for workers.

In Figure 11.15 the initial demand curve for workers (without the fringe) is given by D_1 and the initial supply curve for workers (without the fringe) is given by S_1.[5] Given these curves the market-clearing wage is given by W_1 and the number of workers hired is given by Q_1. Now assume that the employer adds a fringe benefit (say another week of paid vacation each year) that costs exactly the same amount per worker as it is worth to each worker. The demand curve for workers will shift down by an amount equal to the cost per worker of the fringe benefit, or to D2. And the workers supply curve will shift down by the same amount to S2.

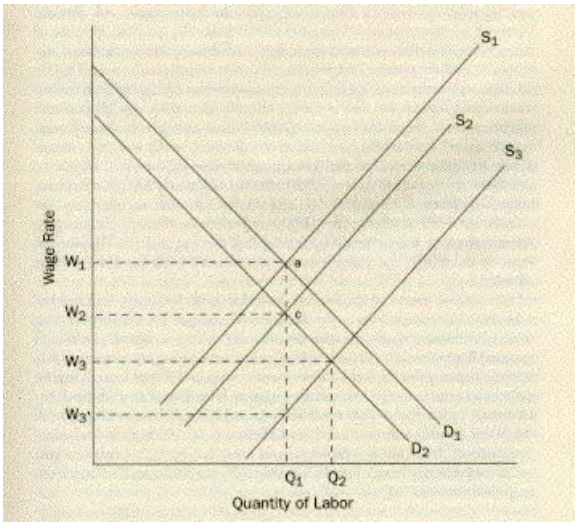


FIGURE 11.15. Fringes and the Labor Market

An Increase in fringes can increase the cost of doing business, causing the demand curve for labor to decrease from *D1* to *D2*. However, it can also cause the supply curve of labor to increase from *S1* to *S3*. The wage rate might fall from W_1 to W_3, but workers are better off because the get the added value of the fringes.

[5] The remaining discussion draws heavily on an article by one of the authors: Dwight R. Lee, "Why Workers Should Want Mandated Benefits to Lower Their Wages," Economic Inquiry (April 1996), pp. 401-407.

These shifts reflect the fact that (assuming workers are worth the same as before and are just as willing to work as before) once the additional vacation is provided: 1) the employer is willing to hire the same number of workers as before if the cost for each remains the same, that is if the wage drops by the same amount as the additional vacation cost per worker; and 2) the same number of workers are willing to work if the value of their compensation remains the same, that is if the money wage drops by the same amount as the value each worker receives from the additional vacation. With both the demand and supply curves shifting down by the same amount, they obviously intersect (as shown in Figure 11.15) at the same number of workers, Q_1, but at a wage, W_2, that is less than W_1 by exactly the cost (and value) of the fringe benefit. In essence, nothing has really changed. Workers are receiving compensation that is worth exactly the same as before ($W_2 + ac = W_1$), and the cost of that compensation to the employer is exactly the same.

In the case just examined, it really makes no difference to the workers or to the employer whether the additional vacation is provided or not. So let's forget about additional vacation time and consider a different fringe benefit, say membership in the neighborhood health club, one that cost the employer the same amount to provide, but which is more valuable to the workers. In this case the employer's demand curve for labor remains at D_2. But because of the greater value the workers receive from the health club membership, the supply curve shifts down below S_2, say to S_3, indicating that now people can be enticed into working for the firm at a lower money wage. As seen in Figure 11.15, with the new supply and demand curves, the money wage will decline to W_3 from W_2 and the number of workers hired will increase to Q_2 from Q_1. But the most important thing to notice is that the workers have gained as the money wage is cut. The Q_1 workers who were employed by the firm before receiving the health club membership, value that membership enough that they would continue working even if the money wage fell to W_3'. Even though the money wage is reduced because of the health club membership, each worker is better off by an amount equal to the difference between W_3 and W_3', which we can think of as a bonus. And obviously the $Q_2 - Q1$ newly hired workers are better off since the wage of W_3 and the health club membership are enough for them to voluntarily leave their previous activities. (You can also see that workers are better off by noting that the wage rate of W_3 plus the value of the fringe, the vertical distance between S_1 and S_3, adds to more than W_1.)

It should be clear that the workers would be even better off if instead of a health club membership, the firm found another fringe benefit that would drive their wages down even more. It can be easily shown that if another fringe benefit (for example, flexible scheduling) is more valuable to the workers than the health club membership and that benefit can be provided at the same cost, the money wage will be driven down below W_3. However, again, the workers will be better off (simply because the sum of the lower wage plus the value of the benefit will lead to higher total compensation). The working rule employers should keep in mind: The more valuable the fringe benefit provided for a given cost, the lower the wage but the better off the workers.

It should be clear that employers have a strong motivation to provide fringe benefits that cost less than they are worth to workers. Both employers and employees win when such benefits are provided. Yet many people believe that private businesses are not sufficiently motivated to provide fringe benefits to their workers and that the government should mandate

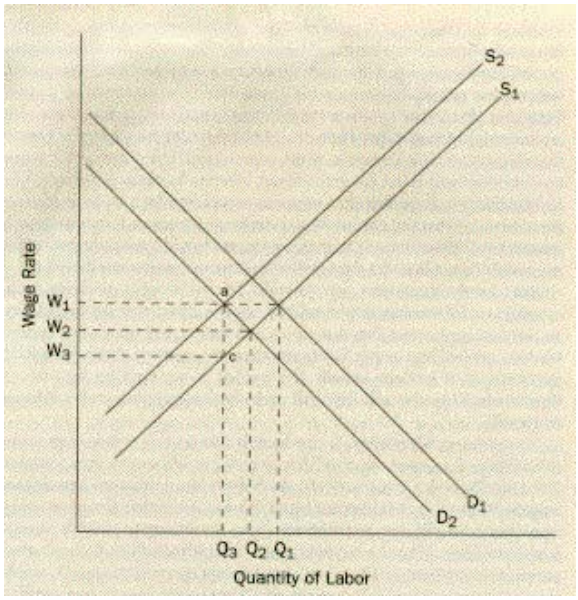
certain employment-related benefits. An explanation for this belief may be the widespread view discussed, which is that workers earn their wages but are given fringe benefits by their employers. This perspective is reflected in the common assumption by advocates of mandated benefits that the cost of those benefits will not result in lower wages for workers.[6] If this were true, then employers would have little motivation to provide fringe benefits even if they were worth more than they cost.

But clearly employers do provide fringe benefits without being required for reasons that should be obvious by now. It pays employers to provide fringe benefits that are worth more than they cost because workers are willing to pay for more than the costs of those benefits in lower wages. If it were true that a fringe benefit, if provided, would not be paid for partly by workers, then it is a fringe benefit that would make both employers and employees worse off.

To see the problem with a fringe benefit that doesn't reduce wages, consider Figure 11.16. Again we start off with a demand and supply curve for labor given by D_1 and S_1 respectively. As before, the initial market-clearing wage is W_1 and Q_1 workers are hired. Now assume that the government mandates a benefit that costs the employer something to provide but which has no value at all to the workers. Such a mandate would shift the demand curve for labor down to D_2, where the vertical distance between D_1 and D_2 is the cost per worker of the benefit, while leaving the supply curve unaffected. As seen in Figure 11.16, the result is a decline in the market-clearing wage to W_2 from W_1 and the layoff of Q_1 - Q_2 workers. Even the workers who keep their jobs are clearly worse off since they end up paying for part of a worthless benefit with lower wages.

FIGURE 11.16 Mandated Benefits and the Labor Market

A mandated benefit that has no value to workers, but imposes a cost of employers, will cause the demand curve to fall from D_1 to D_2. However, the supply curve will not move. A mandated benefit that costs employers but has no impact on wages must be a benefit that his negative value for workers, otherwise, the wage would fall.



[6] This assumption is often made explicit. It is commonly argued, for example, that the one-half of the Social Security tax employers are required by law to pay is really paid by the employer and does not come out of the pocket of the workers.

But what about a mandated benefit that has no effect on wages, one that is paid for entirely by the employer? Such a benefit would be one that had a negative value to workers; it is one that they would be willing to pay to keep from being provided. For example, assume the government mandated that all employers provide a smoke-free work environment. For some employers a smoke-free environment makes sense, and many firms had such a policy before they were required. But consider an employer whose workers all smoke. Providing a smoke-free work place would shift the demand curve down from D_1 to D_2 to reflect the employer cost from workers spending less time working and more time outside smoking. Also, the employer would see the firm's labor supply curve shift back since the workers would find the new working conditions less pleasant than before. If the supply curve shifts back from S_1 to S_2, which is the same amount the demand curve shifts down, then the market-clearing wage remains at W_1 but now $Q_1 - Q_3$ workers are laid off. Even though the wage doesn't fall, the workers are worse off in this case than in any of the previous cases considered -- all of which saw the wage fall. Indeed, the workers are obviously worse off in terms of total compensation: they receive a wage of W_1, but they have to endure the cost of ac from the smoking ban (which means that they receive *net* compensation of W_3 ($W_1 - ac$)

There are at least three important points that follow from the discussion. First, workers benefit from the desire of their employers to cut their wages by providing fringe benefits in much the same way that consumers benefit from suppliers who desire to profit by selling them products. Second, employers have a strong motivation to provide fringe benefits only when those benefits are worth more to workers than they cost. And three, if those who advocate mandated government benefits are correct when they argue that the benefit will be paid for entirely by the employer (will not lower wages), then the benefit isn't worth providing, and mandating it will make workers worse off.

Concluding Comments

Perfect competition is an idealized market structure that can never be fully attained in the real world. Nonetheless, the model helps to illuminate the influence of competition in the marketplace, just as the idealized concepts of the physical sciences help to illustrate the workings of the natural world. Physicists, for example, deal with the concept of gravity by talking of the acceleration of a falling body in a vacuum. Vacuums do not exist naturally in the world as we know it, but as theoretical constructs they are useful in isolating and emphasizing the directional power of gravitational pull. In a similar fashion, the theoretical construct of perfect competition helps to highlight the directional influence and consequences of competition.

The model of perfect competition also provides a benchmark for comparing the relative efficiency of real-world markets. The perfectly competitive model clarifies the rules of efficient production and suggests that free movement of resources is essential to achieving efficient production levels. Without a free flow of resources, new firms cannot move into profitable production lines, increase market supply and push prices down, and force other firms to minimize their production costs. Competition requires mobility of resources.

The model of perfect competition must ultimately be judged not so much by the realism of this underlying assumptions. No "model" is designed to be "real," or fully

descriptive. Rather, they must be judged by their usefulness in understanding behavior. In the manager's corner, we once again used the model of perfect competition to understand why cutting the pay of workers, under some conditions, can be in the interest of workers.

Review Questions

1. In a graph, draw in the short-run average and marginal cost curves, plus the demand curve for a perfect competitor. Give the firm's demanded, identify the short-run production level for a profit-maximizing firm. Identify the profits.
2. On the graph for question 1, indicate with a *Pm* the minimum price the firm requires to continue short-run operations.
3. On the graph for question 1, darken the firm's marginal cost curve above its intersection with the average variable supply cost curve. Explain why that portion of the marginal cost curve is the firm's supply curve.
4. Why does a perfectly competitive firm seek to equate marginal cost with marginal revenue rather than to produce where average total cost is at a minimum?
5. If perfectly competitive firms are making a profit in the short run, what will happen to the industry's equilibrium price and quantity in the long run?
6. Suppose the market demand for a product rises. In the short run, how will a perfect competitor react to the higher market price? Draw a graph to illustrate your answer. What will happen to the market price in the long run? Why?
7. Suppose that you know absolutely nothing about price and cost in a particular competitive industry. How could you nevertheless determine whether the typical firm in the industry was making economic profits or losses?
8. Suppose a manager were to refuse to provide a fringe benefit that could lower the wages of their workers, but were to the benefit of were, on balance. What would be that manager's fate?
9. When should a firm eliminate fringe benefits?

CHAPTER 12

Monopoly Power and Firm Pricing Decisions

That competition is a virtue, at least as far as enterprises are concerned has been a basic article of faith in the American Tradition, and a vigorous antitrust policy has long been regarded as both beneficial and necessary, not only to extend competitive forces into new regions but also to preserve them where they may be flourishing at the moment.

G. Warren Nutter
Henry Alder Einhorn

At the bottom of almost all arguments against the free market is a deep-seated concern about the distorting (some would say corrupting) influence of monopolies. People who are suspicious of the free market fear that too many producers are not controlled by the forces of competition, but instead hold considerable monopoly power. Unless government intervenes, these firms are likely to exploit their power for their own selfish benefit. This theme has been fundamental to the writings of John Kenneth Galbraith.

The initiative in deciding what is produced comes not from the sovereign consumer who, through the market, issues instructions that bend the productive mechanism to his ultimate will. Rather it comes from the great producing organization which reaches forward to control the markets that it is presumed to serve and, beyond, to bend the customers to its needs.[1] Currently, the Department of Justice and nineteen state attorneys general are suing Microsoft because of the concern that one firm has too much "market power." Furthermore, the company, as a consequence, is harming consumers as well as its potential market rivals and may be doing other damage to the economy, for example, impairing competition.

This chapter is really a continuation of our earlier discussion of "market failures," for *monopoly* is often seen as one of the gravest of all forms of failure in markets. Accordingly, we will examine the dynamics of monopoly power and attempt to place their consequences in proper perspective. We will also consider the usefulness of antitrust laws in controlling monopoly and promoting competition. This chapter will elucidate the government's concerns with Microsoft's market position. It will also help us understand Microsoft's court defense. In the next chapter, we will apply the model of monopoly developed here to two forms of partial monopoly, monopolistic competition and oligopoly.

[1] John Kenneth Galbraith, The New Industrial State (Boston: Houghton Mifflin, 1967), p. 6.

The Origins of Monopoly

We have defined the competitive market as the process by which market rivals, each pursuing its own private interests, strive to outdo one another. This competitive market process has many benefits. It enables producers to obtain information about what consumers and other producers are willing to do. It promotes higher production levels, lower prices, and a greater variety of goods and services than would be achieved otherwise.

Monopoly power is the conceptual opposite of competition. Monopoly power is the ability of a firm to raise profitably the market price of its good or service by reducing production. Whereas the demand curve of the competitive firm is horizontal (see the previous chapter), a firm with monopoly power faces a downward-sloping demand curve. By restricting production the monopoly can raise its market price. To maximize its profits (or minimize its losses), such a firm need only search through the various price-quantity combinations. In very general terms, then, a firm with monopoly power is a price searcher. It can control price because other firms are to some extent unable or unwilling to compete. As a result, a monopolized market produces fewer benefits than perfect competition.

Businesses vary considerably in the extent of their monopoly power. The postal service and your local telephone company both have significant monopoly power. They confront few competitors, and entry into their markets is barred by law. IBM has far less monopoly power. Although it can affect the price it charges for its computers by expanding or contracting its sales, IBM is restrained by the possibility that other firms will enter its market. On a smaller scale, grocery stores face the same threat. They may have many competitors already, and they must be concerned about additional stores entering the market. Nevertheless, grocery stores still retain some power to restrict sales and raise their prices.

The exact opposite of perfect competition is pure monopoly. Since, by definition, the pure monopolist is the only producer of a product that has no close substitutes, the demand it confronts is the market demand for the product. Unlike the perfect competitor, who has no power over price, the pure monopolist can raise the price of its product without fear that customers will go elsewhere. With no other producers offering the same product, or even a close substitute, the consumer has nowhere to turn. As we will see, production levels are generally lower and prices higher under pure monopoly than under competition.

How does monopoly arise? To answer that question clearly, we must reflect once again on the basis for competition. Competition occurs because market rivals want to exploit profitable opportunities and can enter markets where such opportunities exist. In the extreme case of perfect competition, there are no barriers to entry, and competitors are numerous. Entrepreneurs are always on the lookout for any opportunity to enter such a market in pursuit of profit. Individual competitors cannot raise their price, for if they do, their rivals may move in, cut prices, and take away all their customers. If a wheat farmer asks more than the market price, for example, customers can move to others who will sell wheat at market price. For this reason perfect competitors are called price takers. They have no real control over the price they charge.

The essential condition for competition is freedom of market entry. In perfect competition entry is assumed to be completely free. Conversely, the essential condition for monopoly is the presence of barriers to entry. Monopolists can manipulate price because such barriers protect them from being undercut by rivals.

Barriers to entry can arise from several sources.

- First, the monopolist may have sole ownership of a strategic resource, such as bauxite (from which aluminum is extracted).
- Second, the monopolist may have a patent or copyright on the product, which prevents other producers from duplicating it. For years, Polaroid had a patent monopoly on the instant-photograph market. (Eastman Kodak developed an alternative process, but was forced to withdraw its camera from the market when a Federal court ruled that it infringed on Polaroid's patent.)
- Third, the monopolist may have an exclusive franchise to sell a given product in a specific geographical area. Consider the exclusive franchise enjoyed by your local telephone company, or was enjoyed, until very recently, your local electric utility.
- Fourth, the monopolist may own the rights to a well-known brand name with a highly loyal group of customers. In that case, the barrier to entry is the costly process of trying to get customers to try a new product.
- Finally, in a monopolized industry, production may be conducted on a very large scale, requiring huge plants and large amounts of equipment. The enormous financial resources needed to produce on such a scale can act as a barrier to entry, because a new entrant operating on a small scale would have costs too high to compete effectively with the dominant firm.

All in all, these external barriers to entry can be thought of as costs that must be borne by potential competitors before they can complete. Such barriers may be "low," which means that a sole producer's monopoly power may be very limited, but such barriers could, theoretically, be prohibitively high.

The Limits of Monopoly Power

Unlike the competitive seller, the monopolist has the power to withhold supplies from the market and to charge more than the competitive market price. Even the pure monopolist's market power is not completely unchecked, however. It is restricted in two important ways. First, without government assistance, the monopolist's control over the market for a product is never complete. Even if a producer has a true monopoly of a good, the consumer can still choose a substitute good whose production is not monopolized. For instance, in most parts of the nation, only one firm is permitted to provide local telephone service. Yet people can communicate in other ways. They can talk directly with one another; they can write letters or send telegrams; they can use their children as messengers. In a more general sense, consumers can use their income to buy rugs or bicycles instead of private lines. To the extent that the individual has alternatives,

his consumption of any good must be considered voluntary. As the Nobel Laureate Friedrich Hayek has written,

> If, for instance, I would very much like to be painted by a famous artist (one who has monopoly power) and if he refuses to paint monopoly efficient for less than a very high price, it would clearly be absurd for monopoly efficient to say that I am coerced. The same is true of any other commodity or service that I can do without. So long as the services of a particular person are not crucial to my existence or the preservation of what I most value, the conditions he exacts for rendering these services cannot be called "coercion."[2]

This is not to say that the effects of monopoly are all positive. If monopoly means that one firm is garnering the assets and markets of all other competitors, it can be viewed as a force that reduces consumer choice. Although the monopolist's coercive power may not be complete, it nevertheless can restrict consumer freedom.

Monopoly power can develop for other reasons. A firm may gain monopoly power because it has built a better mousetrap or developed a good that was previously unavailable. In other words, a firm may be the only producer because it is the first producer, and no one has been able to figure out how to duplicate its product. In this instance, although monopolized, a new product results in an expansion of consumer choice. Furthermore, the monopoly may be only temporary, for other competitors are likely to break into the market eventually.

The monopolist is also restricted by market conditions—that is, by the cost of production and the downward-sloping demand curve for the good. If the monopolistic firm raises its price, it must be prepared to sell less. How much less depends on what substitutes are available. The monopolist must consider as well the costs of expanding production and of trying to prevent competitors from entering the market. The important point here is that there is a range of possible costs and prices at which the monopolistic firm can sell various quantities of a good. Its task is to search through the available price-quantity combinations for the one that maximizes profit.

In a free and open market, monopoly power can be dissolved in the long run. With time, competitors can discover weakly protected avenues through which to invade the monopolist's domain. The Reynolds International Pen Company had a patent monopoly on the first ballpoint pen that it introduced in 1945. Two years later other pen companies had found ways of circumventing the patent and producing a similar but not identical product. The price of ballpoint pens fell from an initial $12.50 to the low prices of today. Many other products that are freely produced today—calculators, video games, car telephones, and cellophane tape, to name a few—were first sold by companies that enjoyed short-run monopolies. Thus the imperfection of monopoly power is crucial. In the long run, excessively high prices, restricted supply, and high profits give potential competitors the incentive to find and exploit imperfections in the monopolist's power. Like the proverbial hole in the dike, those imperfections can undermine even the strongest barrier.

---

[2] F.A Hayek, The Constitution of Liberty (Chicago: University of Chicago Press, 1960). P. 136.

One of the most effective ways for a monopoly to retain its market power is to enlist the coercive power of the state in protecting or extending it boundaries. This strategy has been used effectively for decades in the electric utilities industry and the cable television market. The insurance industry and the medical profession, both of which are protected from competition through licensing procedures, are also good examples. Even the power of the state may not be enough to shield an industry from competition forever. Consumer tastes and the technology of production and delivery can change dramatically over the very long run. The franchise-monopoly of electric power companies, for example, is slowly being weakened by the introduction of home solar power. The railroad industry's market, which was protected from price competition by the state for almost a century, has been gradually eroded by the emergence of new competitors, principally airlines, buses, and trucks. Even the first-class mail monopoly of the U.S. Postal Service is being eroded by Federal Express and other overnight delivery firms. In the long run, government protection may be extended to the very competitors who arise to break a state-protected monopoly. (Such was the case, until recently, with the airline, bus, and tracking industries.)

Should government attempt to break up all monopolies? Since without state protection monopoly may eventually dissipate, the relevant public policy questions are how long the monopoly power is likely to persist if left alone, and how costly it will be while it lasts, in terms of lost efficiency and unequal distribution of income. The machinery of government needed to dissolve monopoly power is costly in itself. Thus the decision whether to prosecute antitrust violations depends in part on the costs and benefits of such an action. Often the rise of a monopoly does warrant government action, but in some cases the benefits of action cannot justify the costs. As described in Chapter 3, the first seller of land calculators enjoyed a temporary monopoly of the U.S. market in 1969. Subsequently the industry developed very rapidly, however, and in retrospect it is clear that a long, drawn-out antitrust action would have been inappropriate.

To give another example, in 1969 the Justice Department found that IBM enjoyed an unwarranted monopoly of the domestic computer market, which was dominated by large mainframe computers. It concluded that an antitrust suit against IBM was justified. Prosecution of the case, which the Justice Department dropped in January 1982, took more than a decade. The accumulated documentation from the proceedings filled a warehouse, and the Justice Department and IBM devoted an untold number of lawyer-hours to the case. In the meantime, IBM's alleged monopoly was seriously eroded by new firms producing mini-and microcomputers, a trend that has continued (and accelerated) since 1982. Thus the net benefits to society from the antitrust action against IBM are at best debatable, and probably negative. That is, the costs most likely exceeded the benefits.

Equating Marginal Cost with Marginal Revenue

In deciding how many times a week to play tennis, an athlete weights the estimated benefits of each game against its costs. Producers of goods follow a similar procedure, although the benefits of production are measured in terms of revenue acquired rather than personal utility. A producer will produce another unit of a good if the additional (or

marginal) revenue it brings is greater than the additional cost of its production—in other words, if it increases the firm's profits. The firm will therefore expand production to the point where marginal cost equals marginal revenue (MC = MR). This is a fundamental rule that all profit-maximizing firms follow, and monopolies are no exception.

Suppose you are in the yo-yo business. You have a patent on edible yo-yos, which come in three flavors—vanilla, chocolate, and strawberry. (We will assume there is a demand for these products.) The cost of producing the first yo-yo is $0.50, but you can sell it for $0.75. Your profit on that unit is therefore $0.25 ($0.75 - $0.50). If the second unit costs you $0.60 to make (assuming increasing marginal cost) and you can sell it for $0.75, your profit for two yo-yos is $0.40 ($0.25 profit on the first plus $0.15 profit on the second). If you intend to maximize your profits, you—like the perfect competitor—will continue to expand production until the gap between marginal revenue and marginal cost disappears. As a monopolist, however, you will find that your marginal revenue does not remain constant. Instead, it falls over the range of production.

The monopolist's marginal revenue declines as output rises because the price must be reduced to entice consumers to buy more. Consider the price schedule in Table 12.1. Price and quantity are inversely related, reflecting the assumption that a monopolist faces a downward-sloping demand curve. (Because the monopolist is the only producer of a product, its demand curve is the market demand curve.) As the price falls from $10 to $6 (column 2), the number sold rises from one to five (column 1). If the firm wishes to sell only one yo-yo, it can charge as much as $10. Total revenue at that level of production is then $10. To see more—say, two yo-yos—the monopolist must reduce the price for each to $9. Total revenue the rises to $18 (column 3).

By multiplying columns 1 and 2, we can fill in the rest of column 3. As the price is lowered and the quantity sold rises, total revenue rises from $10 for one unit to $30 for five units. With each unit increase in quantity sold, however, total revenue does not rise by an equal amount. Instead, it rises in declining amounts—first by $10, then $8, $6, $4, and $2. These amounts are the marginal revenue from the sale of each unit (column 4), which the monopolist must compare with the marginal cost of each unit.

At an output level of one yo-yo, marginal revenue equals price, but at every other output level marginal revenue is less than price. Because of the monopolist's downward-sloping demand curve, the second yo-yo cannot be sold unless the price of both units 1 and 2 is reduced from $10 to $9. If we account for the $1 in revenue lost on the first yo-yo in order to sell the second, the net revenue from the second yo-yo is $8 (the selling price of $9 minus the $1 lost on the first yo-yo). For the third yo-yo to be sold, the price on the first two must be reduced by another dollar each. The loss in revenue on them is therefore $2. And the marginal revenue for the third yo-yo is its $8 selling price less the $2 loss on the first two units, or $6.

Thus the monopolist's marginal revenue curve (columns 1 and 4) is derived directly from the market demand curve (columns 1 and 2). Graphically, the marginal revenue curve lies below the demand curve, and its distance from the demand curve

increases as the price falls (see Figure 12.1, above).[3] (More details on the derivation of the marginal revenue curve can be found in the appendix to this chapter.)

TABLE 12.1 The Monopolist's Declining Marginal Revenue

Quantity of Yo-yos Sold (1)	Price of Yo-yos (2)	Total Revenue (col. 1 x col. 2) (3)	Marginal Revenue (change in col. 3) (4)
0	$11	$ 0	$ 0
1	10	10	10
2	9	18	8
3	8	24	6
4	7	28	4
5	6	30	2

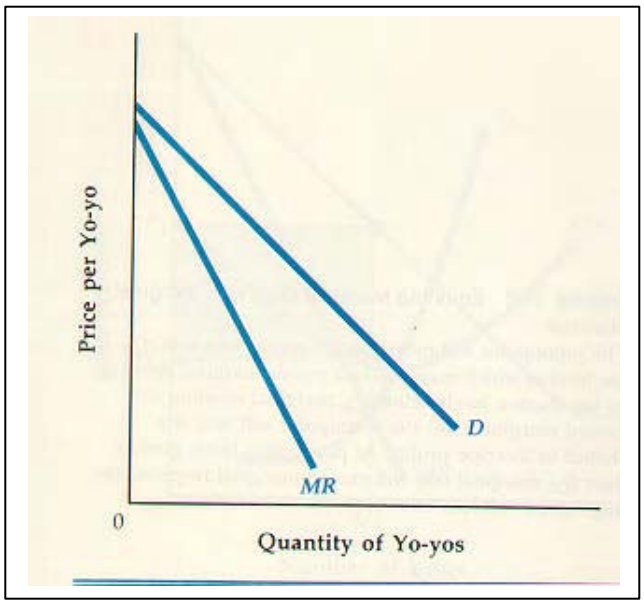


FIGURE 12.1 The Monopolist's Demand and Marginal Revenue Curves

The demand curve facing a monopolist slopes downward, for it is the same as market demand. The monopolist's marginal revenue curve is constructed from the information contained in the demand curve (see Table 12.1).

Figure 12.2 adds the monopolist's marginal cost curve to the demand and marginal revenue curves from Figure 12.1. Because the profit-maximizing monopolist will produce to the point where marginal cost equals marginal revenue, our yo-yo maker will produce Q_2 units. At that quantity, the marginal cost and marginal revenue curves intersect. If the yo-yo maker produces fewer than Q_2 yo-yos -- say Q_1 -- profits are lost

[3] Prove this to yourself by plotting the figures in columns 1 and 2 versus the figures in columns 1 and 4, on a sheet of graph paper. (Another simple way of drawing the marginal revenue curve is to extend the demand curve until it intersects both the vertical and horizontal axes. Then draw the marginal revenue curve starting from the demand curve's point of intersection with the vertical axis to a point midway between the original and the intersection of the demand curve with the horizontal axis. This method can be used for any linear demand curve.)

unnecessarily. The marginal revenue acquired from selling the last yo-yo up to Q_1, MR_1, is greater than the marginal cost of producing it, MC_1. Furthermore, for all units between Q_1 and Q_2, marginal revenue exceed marginal cost. In other words, by expanding production from Q_1 to Q_2, the monopolist can add more to total revenue than to total cost. Up to an output level of Q_2, the firm's profits will rise.

Why does the monopolist produce no more than Q_2? Because the marginal cost of all additional units beyond Q_2 is greater than the marginal revenue they bring. Beyond Q_2 units, profits will fall. If it produces Q_3 yo-yos, for instance, the firm may still make a profit, but not the greatest profit possible. The marginal cost of the last yo-yo up to Q_3 (MC_2) is greater than the marginal revenue received from its sale (MR_2). By producing Q_3 units, the monopolist adds more to cost than to revenues. The result is lower profits.

Once the monopolistic firm selects the output at which to produce, the market price of the good is determined. In this illustration, the price that can be charged for Q_2 yo-yos is P_1. (Remember, the demand curve indicates the price that can be charged for any quantity.) Of all the possible price-quantity combinations on the demand curve, therefore, the monopolist will choose combination a.

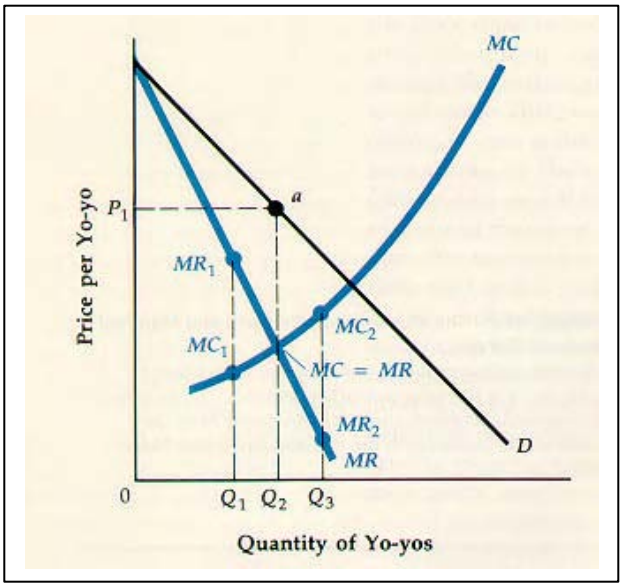


FIGURE 12.2 Equating Marginal Cost with Marginal Revenue

The monopolist will move toward production level Q_2, the level at which marginal cost equals marginal revenue. At production levels below Q_2, marginal revenue will exceed marginal cost; the monopolist will miss the chance to increase profits. At production levels greater than Q_2, marginal cost will exceed marginal revenue; the monopolist will lose money on the extra units. .

Short-Run Profits and Losses

How much profit will a monopolist make by producing where marginal cost equals marginal revenue? The answer can be found by adding the average total cost curve developed in the last chapter to the monopolist's demand and marginal revenue curves (see Figure 12.3). As we have seen, the monopolist will produce where the marginal cost and revenue curves intersect, at Q_1, and will charge what the market will bear for the quantity, P_1. We know also that profit equals total revenue minus total cost (Profit = TR – TC). Total revenue of P_1 times Q_1, or the rectangular area bounded by $0P_1aQ_1$. Total cost is the average total cost, ATC_1, times quantity, Q_1, or the rectangular area bounded by $0ATC_1bQ_1$. Subtracting total cost from total revenue, we find that the monopolist's profit

is equal to the shaded rectangular area ATC_1P_1ab. (Mathematically, the expression profit = P_1Q_1-ATC_1Q_1 can be converted to the simpler form, profit = Q_1 (P_1- ATC_1).)

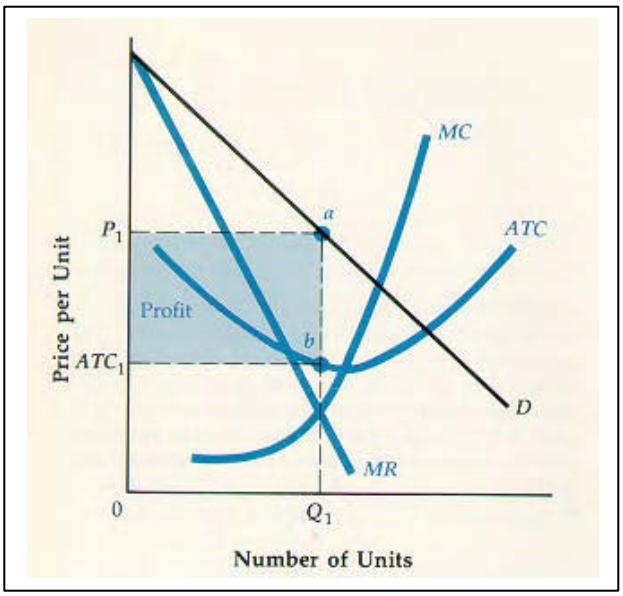


FIGURE 12.3 The Monopolist's Profits

The profit-maximizing monopoly will produce at the level defined by the intersection of the marginal cost and marginal revenue curves: Q_1. It will charge a price of P_1 -- as high as market demand will bear --for that quantity. Since the average total cost of producing Q_1 units is ATC_1, the firm's profit is the shaded area ATC_1P_1ab.

Like perfectly competitive firms, monopolies are not guaranteed a profit. If market demand does not allow them to charge a price that covers the cost of production, they will lose money. Figure 12.4 shows the situation of a monopoly that is losing money. Because losses are negative profits, the monopolist's losses are obtained in the same way as profits, by subtracting total cost from total revenue. The maximum price the monopolist can charge for its profit-maximizing (or in this case, loss-minimizing) output level is P_1, which yields total revenues of P_1Q_1 or $0P_1bQ_1$. Total cost is higher: ATC_1Q_1, or $0ATC_1Q_1$. Thus the monopolist's loss is equal to the shaded rectangular area bounded by P_1ATC_1ab.

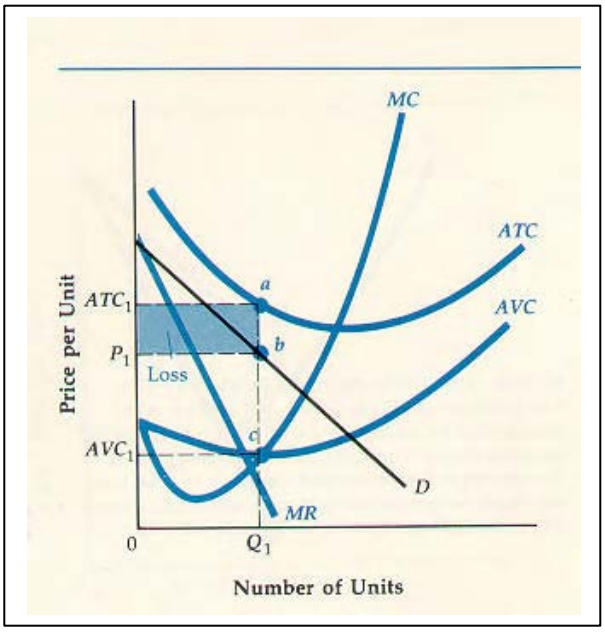


FIGURE 12.4 The Monopolist's Short-Run Losses

Not all monopolists make a profit. With a demand curve that lies below its average total cost curve, this monopoly will minimize its short-run losses by continuing to produce where marginal cost equals marginal revenue (Q_1 units). It will charge P_1, a price that covers its fixed costs, and will sustain short-run losses equal to the shaded area P_1ATC_1ab.

Why does the monopolist not shut down? Because it follows the same rule as the perfect competitor. Both will continue to produce as long as price exceeds average variable cost -- that is, as long as production will help to defray fixed costs. In Figure 12.4, average fixed cost is equal to the difference between average total cost, ATC_1, and average variable cost, AVC_1–or the vertical distance *ac*. Total fixed cost is therefore *ac* times Q_1, or the area bounded by AVC_1 ATC_1ac. Because the firm will suffer a greater loss if it shuts down (AVC_1 ATC_1ac) than if it operates ($P_1.ATC_1ab$), it chooses to operate and minimize its losses.

Of course, in the long run, when the monopoly firm is able to extricate itself from its fixed costs, it will shut down.

Production Over the Long Run

In the long run the monopolistic firm follows the same production rule as in the short run: it equates marginal revenue with long-run marginal cost. In Figure 12.5(a), for instance, the firm produces quantity Q_a, and sells it for price P_a,. (As always, profits are found by comparing the price with the long-run average cost. As an exercise, shade in the profit areas on the figure.) Unlike the perfect competitor, the monopoly firm does not attempt to produce at the lowest point on the long-run average cost cure. With no competition, the monopolistic firm has no need to minimize average total cost. By restricting output, it can charge a higher price and earn greater profits than it can by taking advantage of economies of scale.

Monopolists sometimes do produce at the low point of the long-run average cost curve. They do so only when the marginal revenue curve happens to intersect the long-run marginal and average cost curves at the exact same point [see Figure 12.5(b)] . In this case the monopolist produces quantity Q_b, and sells it at a price of P_b, earning substantial monopoly profits in the process.

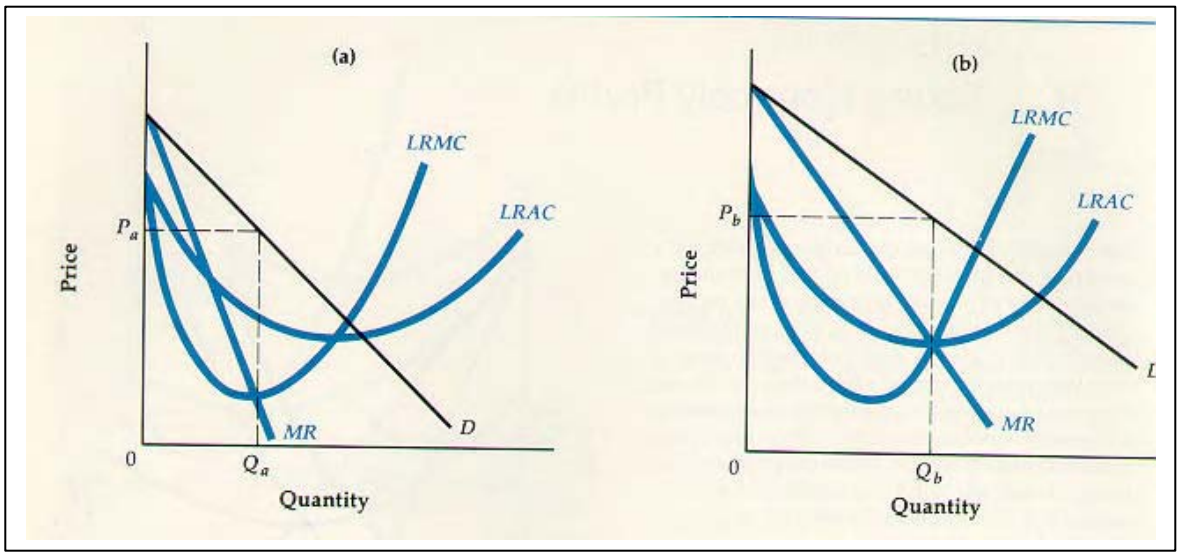
If the demand is great enough, the monopolist will actually produce in the range of diseconomies of scale [see Figure 12.5(c)]. How can the monopolist continue to exist when its price and costs of production are so high? Because barriers to entry protect it from competition. If barriers did not exist, other firms would certainly enter the market and force the monopolistic firm to lower its price. The net effect of competition would be to induce the monopolist to cut back on production, reducing average production costs in the process.

Monopolists cannot exist without barriers to market entry. If other firms had access to the market, the monopolist's profit would be its own undoing—for profit is what others want and will seek, if they can enter the market.

The Comparative Inefficiency of Monopoly

The last chapter concluded that in a perfectly competitive market, firms tend to produce at the intersection of the market supply and demand curves. That point (*b* in Figure 12.6) is the most efficient production level, in the sense that the marginal benefit to the

consumer of the last unit produced equals its marginal cost to the producer. All units whose marginal benefits exceed their marginal costs are produced. All possible net benefits to the consumer have been extracted from production.



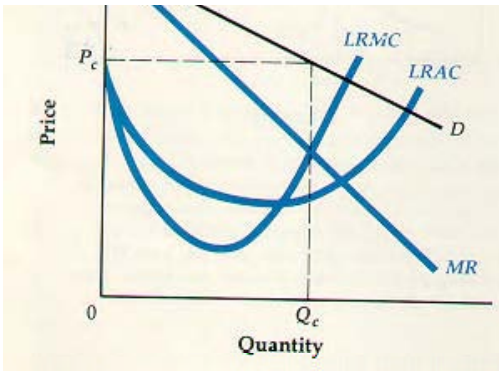


FIGURE 12.5 Monopolistic Production Over the Long Run

In the long run, the monopolist will produce at the intersection of the marginal revenue and long-run marginal cost curves (part (a)] . Unlike the perfect competitor, the monopolist does not bother to minimize long-run average cost by expanding its scale of operation. It can make more profit by restricting production to Q_b and charging price P_b. In part (b), the monopolist produces at the low point of the long-run average cost curve only because that happens to be the point where marginal cost and marginal revenue curves intersect. In part (c), the monopolist produces on a scale beyond the low point of its long-run average cost curve because demand is high enough to justify the cost. In each case, the monopolist charges a price higher than its long-run marginal cost.

When the supply and demand model is applied to a monopolized market, the industry supply curve becomes the monopolist's marginal cost curve (for the monopolist must long-run the plants that in a competitive industry would belong to other producers).[4] Similarly, the industry's demand curve becomes the monopolist's demand. Where as individual competitors must produce at the intersection of supply and demand, the monopolistic firm can choose the price-quantity combination it prefers. By employing fewer resources from production, it can sell a smaller quantity at a higher price. That is just what happens. In short, the monopolist produces less than the competitive level of production -- Q_m instead of Q_c,

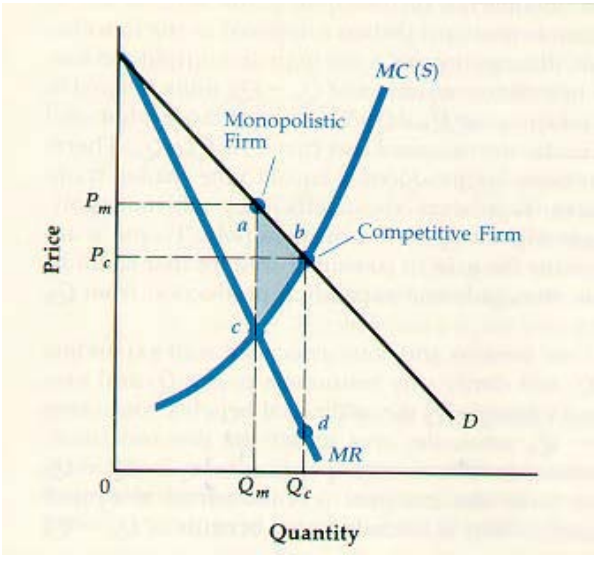


FIGURE 12.6 The Comparative Efficiency of Monopoly and Competition

Firms in a competitive market will tend to produce at point *b*, the intersection of marginal cost and demand (marginal benefit). Monopolists will tend to produce at point *c*, the intersection of marginal cost and marginal revenue, and to charge the highest price the market will bear -- P_m. In a competitive market, therefore, the price will tend to be lower (P_c) and the quantity produced greater (Q_c) than in a monopolistic market. The inefficiency of monopoly is shown by the shaded triangular area *abc*, the amount by which the benefits of producing $Q_c - Q_m$ units (shown by the demand curve) exceed their marginal cost of production.

For each unit between Q_m and Q_c, the marginal benefits to the consumer, as illustrated by the market demand curve, are greater than the marginal costs of production. These are net benefits that consumers would like to have, but that are not delivered by the monopolistic firm interested in maximizing profits, not consumer welfare. The resources that are not used for their manufacture must either remain idle or be used in a less valuable line of production. (Remember, the cost of doing anything is the value of the next-best alternative forgone.) In this sense, economists say that resources are misallocated by monopoly. Too few resources are used in the monopolistic industry, and too many elsewhere.

On balance, then, the inefficiency of monopoly is the benefits lost to consumers when production is restricted. When compared to the outcome under perfect competition, monopoly price is too high and output too low. In Figure 12.6, the gross benefit to consumers of $Q_c - Q_m$ units is equal to the area under the demand curve, or $Q_m abQ_c$. The cost of those additional units is equal under the marginal cost curve, or $Q_m cbQ_c$. Therefore the net benefit of the units not produced is equal to the shaded triangular area *abc*. This area represents the inefficiency of monopoly, sometimes called the dead-weight welfare loss of monopoly. To put it another way, area *abc* represents the gain in

[4] The industry supply curve is not the monopolist's supply curve, however, for a firm's supply is its price-quantity relationship—and a monopolist's price will always exceed its marginal cost.

consumer welfare that could be achieved by dissolving the monopoly and expanding production from Q_m to Q_C. This area helps explain consumers prefer Q_C and producers prefer Q_m. Figure 12.7(a) shows the additional benefits consumers would receive from $Q_c - Q_m$ units, the area under the demand curve, $Q_m abQ_c$. The additional money consumers must pay producers for $Q_c - Q_m$ units, shown by the area under the marginal revenue curve, is a much smaller amount: only $Q_m cdQ_c$. That is, the additional benefits of $Q_c - Q_m$ units, exceed the cost to consumers by the shaded area *abcd*. Consumers obviously gain from an increase in production.

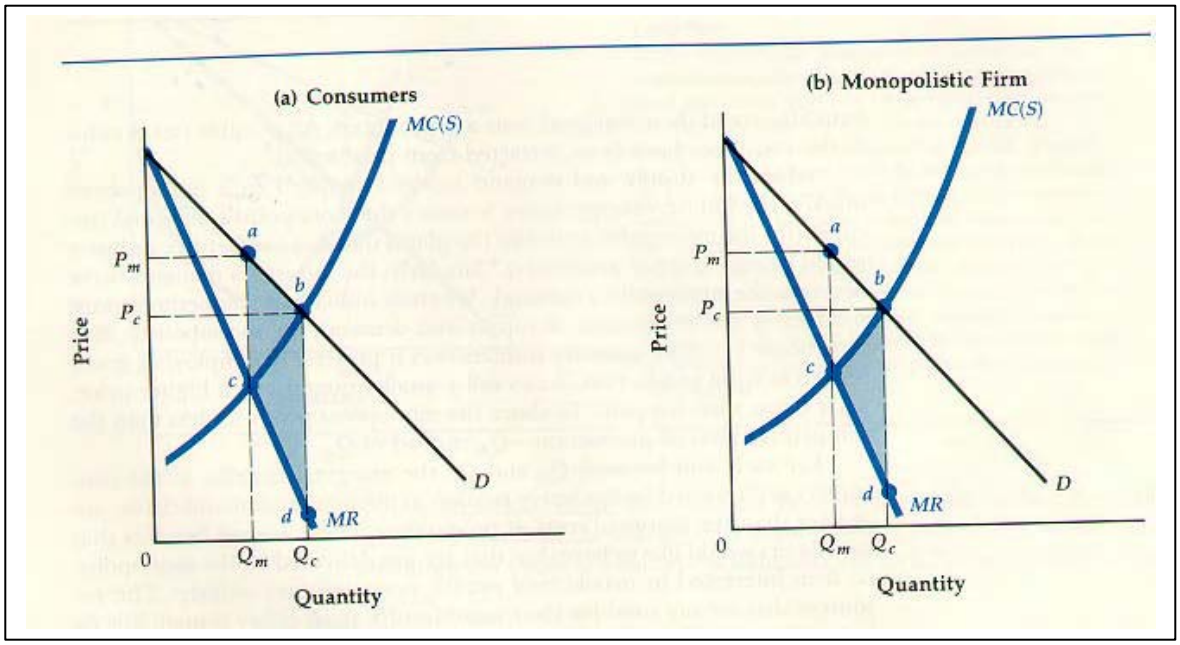


FIGURE 12.7 The Costs and Benefits of Expanded Production

If the monopolist expands production from Q_m to Q_c in part (a), consumers will receive additional benefits equal to the area bounded by $Q_m abQ_c$. They will pay an amount equal to the area $Q_m cdQ_c$ for those benefits, leaving a net benefit equal to the vertically striped area *abcd*. To expand production, the monopoly must incur additional production costs equal to the area $Q_m cbQ_c$ in part (b). It gains additional revenues equal to the area $Q_m cdQ_c$, leaving a net loss equal to the shaded area *cbd*. Thus expanded production helps the consumer but hurts the monopolist.

Yet for virtually the same reason, the monopolistic firm is not interested in providing $Q_c - Q_m$ units. It must incur an additional cost equal to the area $Q_m cdQ_c$ (part (b)], while it can expect to receive only $Q_m cdQ_c$ in additional revenues. The extra cost incurred by expanding production from Q_m to Q_c exceeds the additional revenue acquired by the horizontally stripped area *cbd*. Thus an increase in production will reduce the monopolistic firm's profits (or increase its losses). Notice that consumers would gain more from an increase in production than the monopolist would lose. The shaded area in part (a) is larger than the shaded area in part (b). The difference is the triangular area *abc*.

Price Discrimination

A grocery store may advertise that it will sell one can of beans for $0.30, but two cans for $0.55. Is the store trying to give customers a break? Sometimes this kind of pricing may simply mean that the cost of producing additional cans decreases as more are sold. At other times it may indicate that customer's demand curves for beans are downward sloping and that the store can make more profits by offering customers a volume discount than by selling beans at a constant price. In other words, the store may be exploiting its limited monopoly power.

Consider Figure 12.8. Suppose the demand curve represents your demand for beans, and the supply curve represents the store's marginal cost of producing and offering the beans for sale. If the store charges the same price for each can of beans, it will have to offer them at $0.25 each to induce you to buy two. Its total revenues will be $0.50. As the graph shows, however, you are actually willing to pay more for the first can -- $0.30—than for the second. If the store offers one can for $0.30 and two cans for $0.55, you will still buy two cans, but its revenues from the sale will be $0.55 instead of $0.50.[5] Similarly, to entice you to buy three cans, the store need only offer to sell one for $0.30, two for $0.55, and three for $0.75, and its profits will rise further.[6] The deal does not change the marginal cost of providing each can, which is below the selling price for the first two units and equal to the selling price for the third. The marginal cost of the first can is $0.09; the second, $0.14; and the third, $0.20. The total cost of the three cans to the store is $0.43, regardless of how the cans are priced.

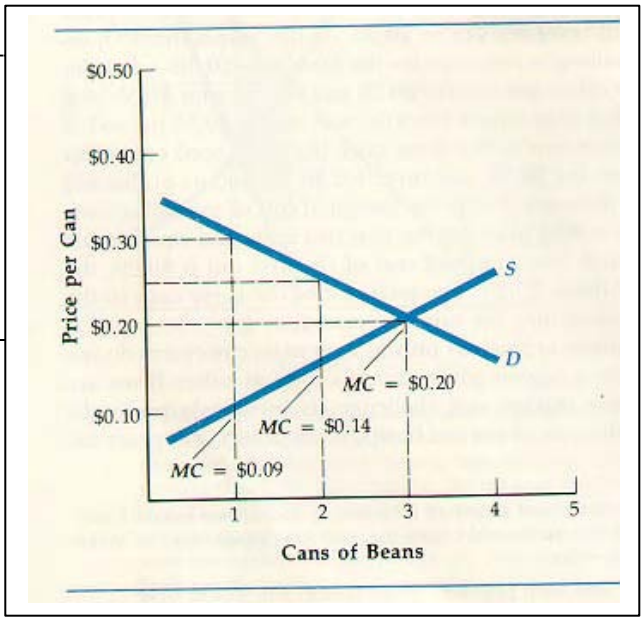


FIGURE 12.8 Price Discrimination

By offering customers one can of beans for $0.30, two cans for $0.55, and three cans for $0.75, a grocery store collects more revenues than if it offers three cans for $0.20 each. In either case, the consumer buys three cans. But by making the special offer, the store earns $0.15 more in revenues per customer.

[5] Notice that if the store had tried to sell all its beans at $0.30, you would have bought only one can, and the store would have forgone the opportunity to make a profit on the second can. Why?

[6] Notice that if the cans had been priced at $0.25 apiece, you would have purchased only two cans. Can you explain the apparent contradiction?

A firm can discriminate in this way only as long as its customers do not resell what they buy for a higher price—and as long as other firms are unable to move into the market and challenge its monopoly power by lowering the price. In the case of canned beans, resale is not very practical. The person who buys three cans has little incentive to seek out someone who is willing to pay $0.25 instead of $0.20 for one can. The profit potential—five cents—is just not great enough to bother with. Suppose a car dealer has two identical automobiles carrying a book price of $5,000 each, however. If the dealer offers one car for $5,000 and two cars for $9,000, many people would be willing to buy both cars and spend the time needed to find a buyer for one of them at $4,500. The $500 gain they stand to make would compensate them for their time and effort in searching out a resale.

Thus advertised price discrimination is much more frequently found in grocery stores than in car dealerships. **Price discrimination** is the practice of varying the price of a given good or service according to how much is bought and who buys it, supposing that marginal costs do not differ across buyers. Car dealers also discriminate with regard to price, however. The salesperson who in casual conversation asks a customer's age, income, place of work, and so forth is actually trying to figure out the customer's demand curve, so as to get as high a price as possible. Similarly, many doctors and lawyers quietly adjust their fees to fit their clients' incomes, using information they obtain from client questionnaires. Whether price discrimination is unadvertised and based on income, as in the case of doctors and car dealers, or advertised and based on volume sold, as in the case of utilities and long-distance phone companies, the important point is that the products or services involved are typically difficult if not impossible to resell.

Some monopolies' products are not difficult to resell, and so they cannot engage in price discrimination. For example, copyright law gives the publishers of economics textbooks some monopoly power, but textbooks are easily resold, both through a network of used-book dealers and among students. Thus, although textbook publishers can alter their sales by changing the price, they rarely engage in price discrimination. Nor do they encourage college bookstores to price-discriminate in their sales to students. The discounts publishers give bookstores on large sales reflect cost differences in handling large and small orders, not students' or professors' downward-sloping demand curves for books. The same can be said about a host of other products protected by patents and copyrights.

The monopolist whose production level was shown in Figure 12.6 could not discriminate among buyers or units bought by each buyer. A monopolist who has such power, however, can produce at a higher output level than Q_m and earn greater profits. Just how much greater depends on how free, or "perfect," the monopolist's power to discriminate is.

Perfect Price Discrimination

The monopolist represented Figure 12.9 can charge a different price for each and every unit sold. Theoretically, this firm has the power of perfect price discrimination ("perfect" from the standpoint of the producer, not the consumer). **Perfect price**

discrimination is the practice of selling each unit of a given good or service for the maximum possible price. Under perfect price discrimination, the seller's marginal revenue curve is identical to the seller's demand curve. In Figure 12.10, for instance, the firm's marginal revenue curve is not separate and distinct from its demand curve, as in Figure 12.7. Its demand curve is its marginal revenue curve. If the first unit can be sold for a price of, say, $20, the marginal revenue from that unit is equal to the price, $20. If the next unit can be sold for $19.95, the marginal revenue from that unit is again the same as the price; and so on. In short, the seller extracts the entire consumer surplus.

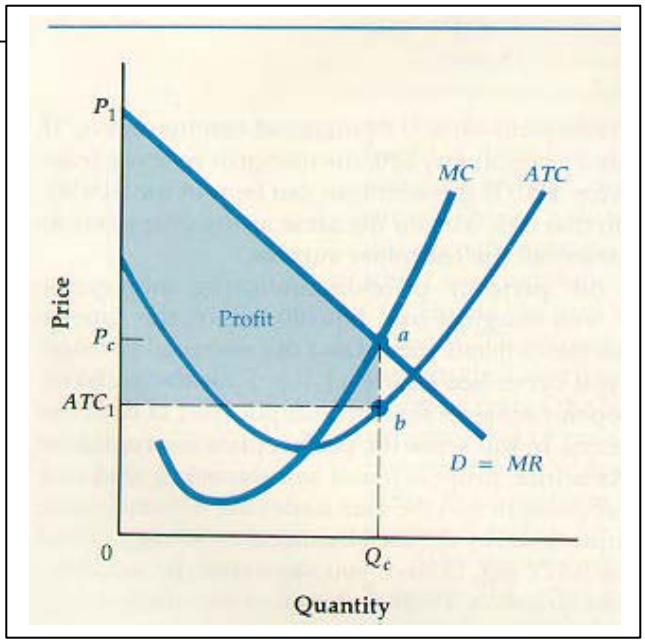


FIGURE 12.9 Perfect Price Discrimination

The perfect price-discriminating monopolist will produce where marginal cost and marginal revenue are equal (point *a*). Its output level, Q_c is therefore the same as that achieved under perfect competition. But because the monopolist charges as much as the market will bear for each unit, its profits—the shaded area ATC_1P_1ab—are higher than the competitive firm's.

As in Figure 12.3, the perfectly price-discriminating monopolist equates marginal revenue with marginal cost. Equality occurs this time at point *a*, the intersection of the demand curve (also the marginal revenue curve) with the marginal cost curve (see Figure 12.9). Thus the perfectly price-discriminating monopolist achieves the same output level as does the perfectly competitive industry. In this sense the perfect price discriminating firm is an efficient producer. As before, profit is found by subtracting total cost from total revenue. Total revenue here is the area under the demand curve up to the monopolist's output level, or the area bounded by $0P_1aQ_c$. Total cost is the area bounded by $0ATC_1bQ_c$ (found, you may recall, by multiplying average total cost times quantity). Profit is therefore the shaded area above the average total cost line and below the demand curve, bounded by ATC_1P_1ab.

Through price discrimination the monopolist increases profits (compare Figure 12.3). Consumers also get more of what they want, although not necessarily at the price they want. In the strict economic sense, perfect price discrimination increases the efficiency of a monopolized industry. Consumers would be still better off if they could pay one constant price, P_c, for the quantity Q_c, as they would under perfect competition. This, however, is a choice the price-discriminating monopolist does not allow.

Discrimination by Market Segment

Charging a different price for each and every unit sold to each and every buyer is of course improbable, if not impossible. The best most producers can do is to engage in imperfect price discrimination—that is, to charge a few different prices, like the grocery store that sells beans at different rates. **Imperfect price discrimination** is the practice of charging a few different prices for different consumption levels or different market segments (based on location, age, income, or some other identifiable characteristic that is unrelated to cost differences). The practice is fairly common. Electric power and telephone companies engage in imperfect price discrimination when they charge different rates for different levels of use, measured in watts or minutes. Universities try to do the same when they charge more for the first course taken than for any additional course. Both practices are examples of multipart price discrimination. Drugstores price-discriminate when they give discounts to senior citizens and students, and theaters price discriminate by charging children less than adults. In those cases, discrimination is based on market segment—namely, age group. By treating different market segments as having distinctly different demand curves, the firm with monopoly power can charge different prices in each market.

Figure 12.10 shows how discrimination by market segment works. Two submarkets, each with its own demand curve, are represented in parts (a) and (b). Each also has its own marginal revenue curve. To price its product, the firm must first decide on its output level. To do so it adds its two marginal revenue curves horizontally. The combined marginal revenue curve it obtains is shown in part (c) of the figure. The firm must then equate this aggregate marginal revenue curve with its marginal cost of production, which is accomplished at the output level Q_m in part (c).

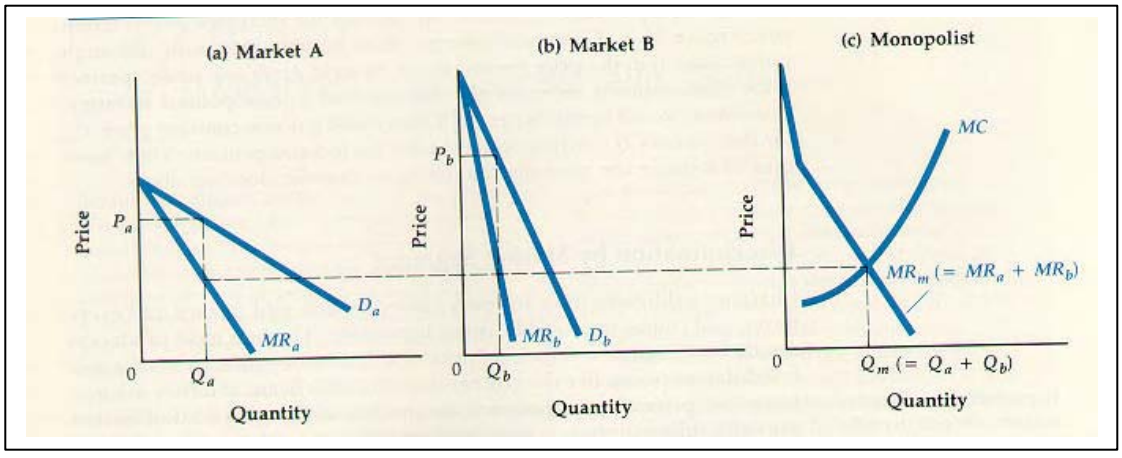


FIGURE 12.10 Imperfect Price Discrimination

The monopolist that cannot perfectly price-discriminate may elect to charge a few different prices by segmenting its market. To do so, it divides its market by income, location, or some other factor and finds the demand and marginal revenue curves in each (part (a) and (b)] . Then it adds those marginal revenue curves horizontally to obtain its combined marginal revenue curve for all market segments, MR_m (part (c)] . By equating marginal revenue with marginal cost, it selects its output level, Q_m. Then it divides that

quantity between the two market segments by equating the marginal cost of the last unit produced (part (c)] with marginal revenue in each market (Parts (a) and (b)] . It sells Q_a in market A and Q_b in market B, and charges different prices in each segment. Generally, the price will be higher in the market segment with the less elastic demand (part (b)] .

Finally, the firm must divide the resulting output, Q_m, between markets A and B. The division that maximizes the firm's profits is found by equating the marginal revenue in each market (shown in parts (a) and (b)] with the marginal cost of the last unit produced (part c). That is, the firm equates the marginal cost of producing the last unit of Q_m, (part (c)] with the marginal revenue from the last unit sold in each market segment ($MC = MR_a = MR_b$). For maximum profits, then, output Q_m, must be divided into Q_a for market A and Q_b for market B.

Why does selling where $MC = MR_a = MR_b$ result in maximum profit? Suppose MR_a were greater than MR_b. Then by selling one more unit in market A and one less unit in market B, the firm could increase its revenues. Thus the profit-maximizing firm can be expected to shift sales to market A from market B until the marginal revenue of the last unit sold in A exactly equals the marginal revenue of the last unit sold in B.

Having established the output level for each market segment, the firm will charge whatever price each segment will bear. In market A, quantity Q_a will bring a price of P_a. In market B, quantity Q_b will bring P_b. (Note that the price-discriminating monopolist charges a higher price in a market with the less elastic demand—market B.) To find total profit, add the revenue collected in each market segment (parts (a) and (b)] and subtract the total variable cost of production (the area under the marginal cost curve in part (c)] and the fixed cost.

Applications of Monopoly Theory

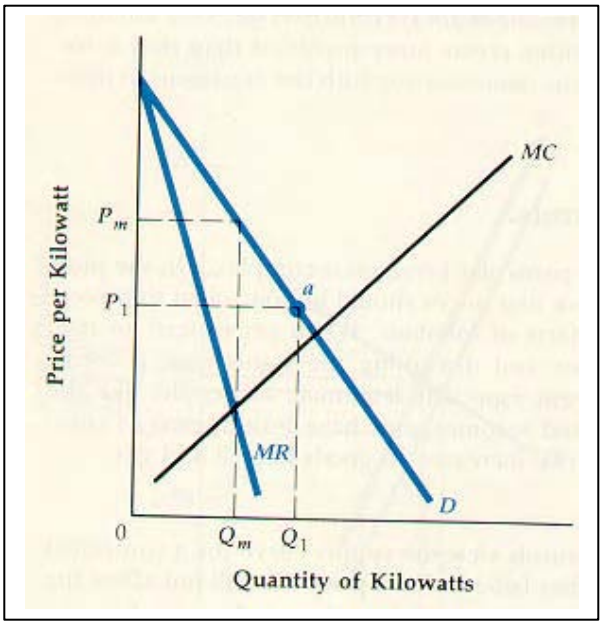
Economics is a fascinating course of study because it often leads to counterintuitive conclusions. This is clearly the case with monopoly theory, as we can show with several policy issues relating to monopoly.

Price Controls under Monopoly

Market theory suggests that price controls can cause monopolistic firms to increase their output. Figure 12.11 shows the pricing and production of a monopolistic electric utility. Without price controls, a firm with monopoly power will produce Q_m kilowatts and sell them at P_m. If the government declares that price too high, it can force the firm to sell at a lower price—for example, P_1. At that price the firm can sell as many as Q_1 kilowatts. With the price controlled at P_1, the firm's marginal revenue curve for Q_1 units becomes horizontal at P_1a. Every time it sells an additional kilowatt, its total revenues will rise by P_1.

As we stressed in the last chapter, the firm's ideal production level is the point at which marginal cost equals marginal revenue. If the firm cannot exactly equate marginal

cost and marginal revenue, it should strive to come as close as possible. With the maximum price controlled at P_1, the firm can increase its revenues by selling up to Q_1 units, which is all demand will permit. At that point marginal revenue approaches but does not equal marginal cost (MC). (To equate marginal revenue with marginal cost, the firm would have to expand production past Q_1, the limit consumers will buy.) Notice that Q_1 is greater than Q_m, the amount the firm would produce under a free but monopolized market. In short, price controls can cause a firm with market power to expand production. (Some exceptions to this rule will be described later.)



FIGUREI2.11 The Effect of Price Controls on the Monopolistic Production Decision

In a free market, a monopolistic utility will produce Q_m kilowatts and will sell them for P_m. If the firm's price is controlled at P_1, however, its marginal revenue curve will become horizontal at P_1. The firm will produce Q_1 -- more than the amount it would normally produce.

Taxing Monopoly Profits

Some people claim that the economic profits of monopoly can be taxed with no loss in economic efficiency. By definition, economic profit represents a reward to the resources in a monopolized industry that is greater than necessary to keep those resources employed where they are. It also represents a transfer of income, from consumers to the owners of the monopoly. Therefore a tax extracted solely from the economic profits of monopoly should not affect the distribution of resources and should fall exclusively on monopoly owners—so the argument goes.

Figure 12.12 shows the reasoning behind this position. This monopoly produces Qm_1, charges Pm_1, and makes an economic profit equal to the shaded area $ATC_1P_{m1}ab$. Since marginal cost and marginal revenue are equal at Q_{m1}, the firm is earning its maximum possible profit. Expansion or contraction of production will not increase its profit. Even if the government were to take away 25, 50, or 90 percent of its economic profit, then the firm would not change its production plans or its price. Nor would it raise prices to pass the profits tax on to consumers. The monopolist price-quantity combination, Pm_1 and Qm_1, leaves the monopolist with the largest after-tax profit—regardless of the tax rate.

The economic profit shown on the graph is not the same as the firm's book profit, however. Book profit tends to exceed economic profit by the sum of the owners' opportunity cost and risk cost. For practical reasons, government must impose its tax on

book profit, not economic profit. As a result, the tax falls partly on the legitimate costs of doing business, shifting the firm's marginal cost curve upward, from MC_1 to MC_2 in Figure 12.12 The monopolist, in turn, will reduce the quantity produced from Q_{m1} to Q_{m2}, and raise the price from P_{m1} to P_{m2}. Thus part of the government tax on profits is passed along to consumers as a price increase. Consumers are doubly penalized—first through the monopoly price, which exceeds the competitive price, and second through the surcharge added by the profits tax.

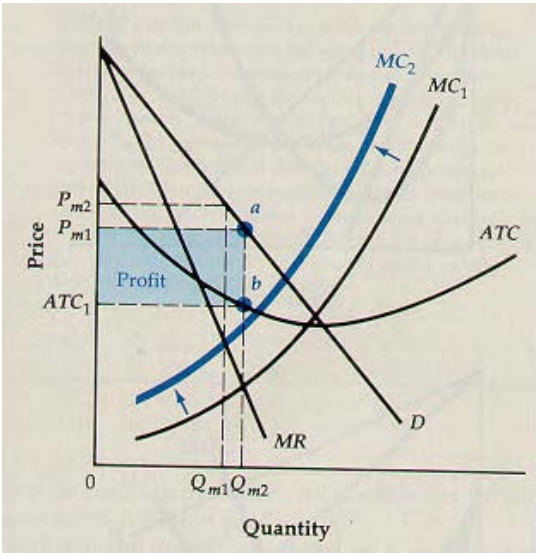


FIGURE 12.12 Taxing Monopoly Profits

Theoretically, a tax on the economic profit of monopoly will not be passed on to the consumer—but taxes are levied on book profit, not economic profit. As a result, a tax shifts the firm's marginal cost curve up, from MC_1 to MC_2, raising the price to the consumer and lowering the production level.

Monopolies in "Goods" and "Bads"

Because monopolies restrict output, raise prices, and misallocate resources, students and policy-makers tend to view them as market failures that should be corrected by antitrust action. If a monopolized product or service represents an economic good—something that gives consumers positive utility—restricted sales will necessarily mean a loss in welfare.

Some products and services, however, may be viewed as "bads" by large portions of the citizenry. Drugs, prostitution, contract murder, and pornography may be goods to their buyers, but they represent negative utility to others in the community. Thus monopolies in the production of such goods may be socially desirable. If a drug monopoly attempts to increase its profits by holding the supply of drugs below competitive levels, most citizens would probably consider themselves better off.

The question is not quite that simple, however. A heroin monopoly may restrict the sale of heroin in a given market. Yet because the demand for heroin is highly inelastic (because of drug addiction), higher prices may only increase buyers' expenditures, raising the number of crimes they must commit to support their habit. Paradoxically then, reducing heroin sales could lead to more burglaries, muggings, and bank hold-ups.

Of course, drugs and other underground services are not normally subject to antitrust action; they are illegal. The analogy may be applied to legal goods and services, however, such as liquor. Given the negative consequences of drinking, as well as religious prohibitions, many people might consider alcoholic beverages an economic bad.

In that case a state-long-run liquor monopoly could provide a social service. By restricting liquor sales through monopoly pricing, it would reduce drunk driving, thus limiting the external costs associated with drinking. (The same objective—fewer liquor sales and less drunk driving—could also be accomplished through higher taxes.)

The Total Cost of Monopoly

High prices and restricted production are not the only costs of monopoly. The total social cost of monopoly power is actually greater than is shown by the supply and demand model in Figure 12.6. Many firms attempt to achieve the benefits of monopoly power by erecting barriers to entry in their markets. The resources invested in building barriers are diverted from the production of other goods, which could benefit consumers. The total social cost of monopoly should also include the time and effort that the antitrust Division of the Department of Justice, the Federal Trade Commission, state attorneys general, and various harmed private parties devote to thwarting such attempts to gain monopoly power and to breaking it up when it is acquired.

Another, subtler social cost of monopoly is its redistributional effect. Because of monopoly power, consumers pay higher prices than under perfect competition (P_m instead of P_c in Figure 12.6). The real purchasing power of consumer incomes is thus decreased, while the incomes of monopoly owners go up. To the extent that monopoly increases the price of a good to consumers and the profits to the producer, then, it may redistribute income from lower-income consumers to higher-income entrepreneurs. Many consider this redistributional effect a socially undesirable one.

In addition, when we measure the inefficiency of monopoly by the triangular area *abc* in Figure 12.6, we are assuming that demand for the monopolized product and all other goods is unaffected by the redistribution of income from consumers to monopoly owners. This may be a reasonable assumption when the monopolist is a maker of vegetable-slicing machines. It is less reasonable for other monopolies, such as the postal, local telephone, and electric power services. Those firms, which are quite large in relation to the entire economy, can shift the demand for a large number of products, causing further misallocation of resources.

Finally, our analysis has assumed that a monopoly will seek to minimize its cost structure, just as perfect competitors do. That may not be a realistic assumption because the monopolist does not, by definition, face competitive pressure. If a monopoly relaxes its attentiveness to costs, the result can be the inefficient employment of resources.

Why a Durable Goods Monopoly Must Charge the Competitive Price

If prohibitive barriers to entry protect it, can a monopolist always charge the monopoly price indicated? University of Chicago Professor of Law and Economics and Noble Laureate Ronald Coase wrote a very famous article years ago in which he pointed out

that even a monopolistic producer of a durable good would charge a competitive price for its product.[7]

Why? Because no sane person would buy all or any portion of the durable good at a price above the competitive level. He used the example of a monopoly owner of a plot of land. If the owner tried to sell the land all at once, he would have to lower the price on each parcel until all the land were bought – where the downward sloping demand for land crossed the fixed vertical supply of land -- which means the owner would have to charge the competitive price (where the demand for the land and the supply of the land came together).

You might think that the sole/monopoly owner of land would be able to restrict sales and get more than the competitive price. However, buyers would reason that the monopoly owner would eventually want to sell the remaining land, but that land could only be sold at less than the price of land already sold. This means that the buyers would rationally wait to buy until the price came down. This means that the owner would sell nothing at the monopoly price, and would only be able to sell the land at the competitive price.

This analysis works out this way only because the land is durable. Monopolies can charge monopoly prices for nondurable goods, and they can do that because they have control over production. This means that one way a monopoly can elevate its price above the competitive level is by somehow making its product less durable. This may explain why many software producers are constantly bringing out new, updated, and upgraded versions of their programs – to, in the minds of consumers, make their programs less than durable.

Still, computer programs must remain, to some degree for some time, "durable," which ultimately imposes a competitive check on dominant software producers, for example, Microsoft. The Justice Department seems to believe that Microsoft doesn't have competitors. Well, and one of Microsoft's biggest competitors is none other than Microsoft itself. Any new version of, say, windows, must compete head to head with the existing stock of old versions, which computer users can continue to use at zero price. That very low price on old versions of Windows imposes a check on the prices that Microsoft can charge on any new version.

Monopolies in Network Goods

The conditions under which monopoly might be expected to emerge and prosper have expanded in recent years with the development of the theory of networks, which we have already introduced. As noted in an earlier chapter, in 1998, the Justice Department filed an antitrust suit against Microsoft for, among other things, engaging in "predatory" pricing in the Internet browser market. The Justice Department argued that by giving away Internet Explorer, Microsoft was attempting to snuff out a serious market rival in

[7] Ronald H. Coase, "Durability and Monopoly," Journal of Law and Economics, vol. 15 (April 1972), pp. 143-149.

the browser market and a potential competitive threat in the operating system market. The Justice Department also argued that the market dominance Microsoft now enjoys in the operating system market could be equated with monopoly power because of the presumed existence of "high switching costs" and "lock-ins."

Switching Costs and Lock-Ins

Once people have adopted the operating system, along with the accompanying computer hardware, and have learned how to use the accompanying applications, there are presumed costs of switching to other operating systems. To switch, people have to buy a different operating system, and maybe different computer equipment, as well as learn new applications that might require different instructions and have a different "look and feel." They might also have to retool and retrain their computer service providers, or switch providers altogether.

Assistant Attorney General Joel Klein introduced "switching costs" into his argument by first repeating his position that Microsoft was convinced that it could not win the browser war based on the relative merits of Internet Explorer. He then quoted Microsoft's Megan Bliss and Rob Bennett, who wrote in an email that the way to increase Internet Explorer's share of the browser market was by "leveraging our strong share of the desktop": "[I]f they get our technology by default on every desk, then they'll be less inclined to purchase a competitive solution. . . ."[8] The Justice Department's chief economist Franklin Fisher gave more details in his testimony for the government, "Where network effects are present, a firm that gains a large share of the market, whether through innovation, marketing skill, historical accident, or any other means, *may* thereby gain monopoly power. This is because it will prove increasingly difficult for other firms to persuade customers to buy their products in the presence of a product that is widely used. The firm with a large market share *may* then be able to charge high prices or slow down innovation without having its business bid away" (emphasis added).[9] Fisher added later, "As a result of scale and network effects, Microsoft's high market share leads to more applications being written for its operating system, which reinforces and increases Microsoft's market share, which in turn leads to still more applications being written for Windows than for other operating systems, and so on."[10]

The government's position on the role of switching costs has been widely adopted in the media. For example, the editors at The Economist have summed up the network effects/switching costs/lock-in line of argument very neatly in their retort to Microsoft's supporters:

---

[8] Joel I. Klein, et. al., Complaint, United States of America vs. Microsoft Corporation, May 20, 1998 (as downloaded from www.usdoj.gov/atr/cases3/micros/1763.htm), p. 38.

[9] Fisher, Direct Testimony, pp. 15-16.

[10] Ibid., p. 27.

> The arguments [that suggest that antitrust laws have no relevance for today's information age], plausible as they may seem, are wrong. 'Network' effects, in which the value of a product depends on the number of users, occur in many high-tech markets – just as they did in earlier industries such as railways and telephones. These effects hugely increase the risk that that one firm may dominate a particular market, probably not forever but certainly for a significant amount of time. True, the products may change, often substantially. But such are the barriers to entry, arising from large installed bases that are locked into a particular technology and from control over distribution, that dominant firm can still remain entrenched.[11]

By suggesting that the operating system market is characterized by network effects that can cause a firm's market share to build on itself, the Justice Department is effectively arguing that Microsoft's current market dominance has been a consequence of economic forces outside of the company's influence. If Microsoft's market position can be viewed as a product of forces of "nature" – or "technology" – then it might be rightfully deduced that Microsoft has itself achieved virtually nothing, which can mean that the Justice Department, by threatening to force Microsoft to put Netscape's icon on the desktop, is not violating any property rights Microsoft may have justly earned.

One of the real problems with the Justice Department's case is that, contrary to the impression left by all the talk about how network effects build on themselves, network effects just don't happen. They are not a part of "nature" or "technology" in the sense that they exist independent of someone (or some firm) causing them to exist. Network effects are truly brought into existence, or are created, as someone (or some firm) works to build the network, and this is necessarily the case. Someone must think of ways to overcome the initial dilemma: How does a network firm get customers to buy the product (operating system) when there are no programs, or how does a firm get program developers to write programs when there are no buyers of the product?

Indeed, the operating system buyers and applications must emerge more or less together, and the emergence process must be coordinated, encouraged, and directed by someone (or some firm). And it should be understood that creating the network is likely to be very expensive, because of buyer and developer resistance, and to require a substantial up-front investment on the part of someone (or some firm) to overcome the resistance – Microsoft, for example.

By arguing that networks are characterized by "high switching costs," the Justice Department is effectively saying that Microsoft's market dominance is protected by an *internal* barrier to entry, which acts like all barriers and restricts entry. Switching costs reduce competition, lower consumer choice, and enable the dominant producer to raise its prices. With high switching costs, the dominant producer doesn't have to worry about its customers switching in response to a higher price, or so the Justice Department argues. Frederick Warren-Bolton, the lead economist for the 19 state attorneys general, reasons

[11] "Lessons from Microsoft," The Economist, March 6, 1999, p. 21.

that computer "users become 'locked in' to a particular operating system [sic]," which implies a barrier to entry and expansion for existing competitors. He adds, "The software 'lock-in' phenomenon creates barriers to entry for new PC operating systems to the extent that consumers' estimate of the switching costs are large relative to the perceived incremental value of the new operating system"[12]

The higher the switching costs, the more the dominant producer can raise its price without fear of the customers switching to existing or new competitors. Indeed, it might be deduced that if switching costs were the only barrier to entry, then a firm's monopoly power – or its ability to raise its price – is limited to the extent of the switching costs. A firm that tries to charge a higher price than that allowed by the switching costs would find that it has left its market open to entry by rivals who would find the consumers perfectly willing to incur the switching costs, because those costs would then be less than the "staying costs" associated with remaining with the established firm.

By introducing the specter of "lock-ins," the Justice Department is seeking to suggest that the switching costs are so high that switching is extremely difficult, if not impossible, thus presumably fortifying its argument that Microsoft has substantial monopoly power. Fisher concedes that "market forces and developments can erode monopoly power based solely on network effects," but that is precisely why, according to Fisher, Microsoft felt compelled to engage in "predatory pricing," to wipe out Netscape as a potential alternative software platform for running personal computers. [13]

Is that the case? Before people accept the Justice Department's arguments, they need, at least, to pause and ask whether Microsoft's pricing strategy is consistent with the dictates of a market entrenched in network effects. If economics of networks dictate zero or below-zero prices, then Microsoft's price for its browser is not necessarily "predatory," contrary to what the Justice Department claims.

Lock-Ins and the QWERTY Keyboard

The path dependency/lock-in theory has gained wide support among many academics and policy makers partially because economic theoreticians and historians have been able to point to two concrete examples of the supposed wrongs of path dependency and lock-ins. The classic, widely cited example of path dependency and lock in is the "QWERTY" keyboard, which takes its name from the way the keys on the far left of the top row of most keyboards line up.

---

[12] Frederick R. Warren-Bolton, Direct Testimony, State of New York *ex rel.* Attorney General Dennis C. Vacco, *et. al.* vs. Microsoft Corporation, Civil Action No. 98-1233 (TPJ), p. 21 (as downloaded from http://www.usdoj.gov/atr/cases/f2000/2079.htm). Warren-Bolton adds, "Often, switching operating systems also means replacing or modifying hardware. Businesses can face even greater switching costs, as they must integrate PCs using the new operating systems and application software within their PC networks and train their employees to use the new software. Accordingly, both personal and corporate consumers are extremely reluctant to change PC operating systems" (Ibid., p. 22).

[13] Fisher, Direct Testimony, p. 16.

According to economic historian Paul David, the arrangement for the keys on this keyboard was first developed in the 1960s for what were then newly invented typewriters, and this arrangement was developed and adopted only because it minimized the prospect for the keys jamming as their arms moved toward the paper.[14] The original keyboard was, supposedly, adopted by one typewriter manufacturer after the other, not because it was potentially the most productive arrangement of keys, but because it was established as the "standard." Manufacturers became further "locked in" to the QWERTY keyboard when touch-typing was developed in the 1880s and then widely taught thereafter. David writes, "The occurrence of this 'lock in' as early as the mid-1890s does appear to have owed something also to the high costs of software 'conversion' and the resulting *quasi-irreversibility of investments* in specific touch-typing skills" (italics in the original).[15]

Because of the "lock in" on the key arrangement that was thought to be a "historical accident," the QWERTY key arrangement is now widely used on computer keyboards, but not because QWERTY is better than all potential alternatives. Indeed, according to this view of keyboard history, "competition in the absence of perfect future markets drove the industry permanently into standardization *on the wrong system* – where decentralized decision making subsequently has sufficed to hold it."[16]

According to what has now become (and proven to be) legend, August Dvorak and W. L. Dealey developed a keyboard (referred to as the Dvorak or DSK keyboard) in 1932 that has, according to David, "long held most of the world's records for speed typing."[17] Moreover, the Navy supposedly showed in experiments that the greater productivity from the Dvorak keyboard could more than cover the cost of the required retraining.[18]

However, the Dvorak keyboard has never gotten a toehold in the keyboard market. Why? The advocates of lock-ins argue there are high switching costs for typists who are used to the QWERTY keyboard; they would have to learn another key arrangement. Typewriter manufacturers have never switched to Dvorak because it did not make good business sense, given they must appeal to the existing typists. Computer keyboard manufacturers adopted the QWERTY key arrangement because they had no other choice, given that all (typewriter) typists, who were potential computer customers, would not buy keyboards with the new key arrangement, in spite of its supposed superiority. The author of the QWERTY story imagined that "there are many more QWERTY worlds [in which an inferior standard is adopted by historical accident] lying

---

[14] As reported by Paul A. David. "Clio and the Economics of QWERTY," American Economic Review, vol. 75 (1985), pp. 332-337.

[15] Ibid., pp. 335-336.

[16] Ibid. p. 336.

[17] Ibid., p. 332.

[18] Ibid., p. 332.

out there in the past, on the very edges of the modern economic analyst's tidy universe; worlds we do not yet fully perceive or understand, but whose influence, like that of dark stars, extends nonetheless to shape the visible orbits of our contemporary economic affairs."[19]

The implication for the Microsoft case is obvious. If the QWERTY story is true, then it is plausible that the operating systems market might be one of those "dark stars" that has become visible, because tens of millions of computer users are similarly locked in to Windows, even though there might be a superior operating system (such as some combination of Netscape's Navigator and Sun's Java programming language) waiting in the wings of modern technology to be adopted. However, the superior system doesn't have a chance of making it in the market because each Windows user does not, by himself or herself, have the requisite incentive to make the switch. Unless large numbers of people make the switch more or less together, then any new user may have a technically superior system that has few applications written for it.

Fortunately for consumers and unfortunately for the Justice Department's case, built partially on the theory of path dependency, the QWERTY story is what we have called it, a legend – a good story that has taken on a life of its own but is not grounded in the facts of keyboard history. University of Texas, Dallas Economics Professor Stan Liebowitz and North Carolina State University Economics Professor Stephen Margolis did what a lot of QWERTY storytellers should have done long ago: they went back and researched the history of keyboards and found that much of the evidence on the supposed superiority of the Dvorak keyboard was from Dvorak's own studies that were poorly designed. Even then, Dvorak's own "evidence was mixed as to whether students, as they progress, retain an advantage when using the Dvorak keyboard since the differences seem to diminish as typing speed increases."[20] The claimed benefits from the Navy study are similarly disputable, and other studies found substantial retraining costs, leading Liebowitz and Margolis to conclude that "the claims for the superiority of the Dvorak keyboard are suspect."[21]

Even if it were proven that the Dvorak keyboard were superior to the QWERTY keyboard, the future gains from making the switch (in present discounted value terms) must be greater than the current costs incurred before it can be said that the "wrong" keyboard continued in use. If the cost of switching were greater than the gains to be gotten from the switch, switching would constitute a net societal loss (as well as a loss for employers and/or typists). Liebowitz and Margolis argue that while David made provocative claims, he never proved his point.

---

[19] Ibid., p. 336.

[20] Stan J. Liebowitz and Stephen E. Margolis. 1990. "The Fable of the Keys," Journal of Law and Economics, 33 (April), 1-25; as reprinted in Stan J. Liebowitz and Stephen E. Margolis, Winners, Losers, and Microsoft: *How Technology Markets Choose Products* (Oakland, Calif.: Independent Institute, 1999), p. II-30 (galley pages).

[21] Ibid., p. II-45.

The Liebowitz/Margolis finding is altogether understandable. If a keyboard were substantially more efficient than the established keyboard, it's hard to see why the new keyboard wouldn't be adopted. Granted, some individual typists might be resistant to making the switch without some outside help. But if the keyboard were substantially superior, then, as we pointed out earlier in this chapter, it follows that the manufacturer should have ample incentive to cover some of the typist's switching costs, through, perhaps, the provision of retraining courses. Companies that hire large numbers of typists, or computer users, would also have ample incentives to buy the new keyboard. They could prorate the retraining costs over a large number of employees from whom they could garner substantial productivity improvements. Their investment in retraining could be expected to have an immediate upward impact on their company's stock price, given that market watchers would expect the productivity improvement to improve the company's long-term profit stream. If the company's executives were not sufficiently wise to make the retraining investment, then surely there would be entrepreneurs outside of the firm who would understand that they could buy control of the company at a low price, change company policies on things like keyboards, and then sell the company at a higher price.

Another similar legend has grown up around how the VHS format for videocassettes tapes and recorders came to dominate the Betamax format, which was supposedly the markedly superior format of the two. The Betamax format may actually be technically superior to the VHS format (we are unwilling to judge), but the VHS format has always had one big advantage over Betamax that counts for more then greater technical attributes such as a clearer pictures: A whole movie could be recorded on a VHS tape, which was not possible on the Betamax. VHS became the adopted format because it met better the needs of the growing home movie rental and sales business.[22]

From both fact and conceptual arguments, the presumption that some combination of path dependency, network effects, switching costs, and lock-ins protect network firms is wrong, or, at the very least, not proven. If the Justice Department wants to claim that Microsoft is protected to the same extent (and for the same reasons) as the QWERTY keyboard and VHS, then we can't help but concur. However, we would insist that our agreement means that Microsoft doesn't have much in the way of long-term market protection.

Surely, we might imagine that there is an operating system out there that is *marginally* superior to Windows and that there might be switching costs that can cause people to resist switching, but it doesn't follow that it always makes sense for people to switch to the superior system (if it is in fact superior and not just claimed to be superior). The superior system would have to be sufficiently superior to the existing system to more than cover the switching costs. Any system that has benefits that (in present value terms) are greater than the switching costs is bound to be adopted, or so it seems to us. Making the switch then makes too much business sense for too many people to expect otherwise. Unlike the classic example of "externalities," in which no one has an incentive to correct

[22]Stan J. Liebowitz and Stephen E. Margolis, "Path Dependence, Lock-In and History," Journal of Law, Economics, and Organization, (1995) vol. 11, pp. 205-226.

the problem, there are market solutions to "network externalities." If Microsoft is acting like a monopolist, then the switch makes even greater business sense, and should be viewed as virtually an irresistible temptation for those that have an economic interest in engineering the switch.

But, then, we have said nothing that is special about the operating system market. Markets for a variety of goods and services have switching costs, and new entrants have to find ways of overcoming such costs. New hamburger restaurants have to overcome customer inertia that might be related to the new restaurant's lack of reputation for good food (and clean restrooms). Banks that wish to operate online have the problem of overcoming people's resistance to doing their banking on a computer. But businesses have been creative in finding new ways to cover the switching costs. New restaurants will often cut their prices below cost, or pass out coupons that have the same effect. A variety of businesses have offered cash payments or discounts for each online transaction made. In 1998, Chase Bank advertised that it would pay online customers $25 for each of the first five online transactions they made. In 1999, after it set up its auction web site, Amazon.com offered book customers a $10 gift certificate on their first auction purchase. If there are efficiency improvements to a switch over to another product that mean greater profits for new firms, "network externalities" (which network effects are sometimes called) may be "external" to buyers, but those network externalities can be "internalized" by entrepreneurial firms. Justice Department action is unnecessary to maintain consumer choice.

MANAGER'S CORNER I: **Getting Prices Right with the Right Incentives**

Incentives are necessarily embedded in a firm's pricing policy. Lowering the price of a product increases the incentive for consumers to purchase. Conversely, increasing the price reduces that incentive. Accordingly, people tend to buy more when the price falls and buy less when the price rises. This inverse relationship between price and quantity is commonly called the "law of demand." For a firm to be successful, it has to choose the "right" price, given the demand (or specifics of the inverse relationship between price and quantity) for its products. The "right" price achieves maximum profit by striking a balance between charging more, selling more, and covering the costs of production.

Saying that the firm must choose the "right" price is easier than actually choosing it. The maxim offers little practical guidance to managers confronting the complex problem of keeping the firm as profitable as possible. For example, managers can never be completely sure what the demand for their company's product is. Moreover, a company's demand is not given from on high; it can be influenced by management decisions. Good managers can increase the demand for their products by improving the quality of those products, increasing the credibility with which those products are advertised and their quality is ensured, and establishing a reputation for honesty and fair dealing. Indeed, much of our previous discussion on different aspects of getting the incentives right can be thought of as aimed at increasing the demand for the goods and

services being produced. But demands are also affected by other factors, and many of them are beyond managers' ability to control or predict.

So managers, no matter how good they are, will always have to make guesses about the demands for their products -- about how much they can sell of their products at different qualities and prices. There are techniques for estimating product demands (a discussion of which goes beyond the purpose of this book) and, though these techniques are never perfect, they can help managers move from making *mere* guesses to making *educated* guesses about their demands.

But even if managers knew all there was to know about the demands for their products, they would still be faced with tough decisions calling for creative pricing strategies. In standard discussions of competitive markets, firms have no choice; they must set the price which competition dictates. Even in the case where the firm has some choice in setting the price there is no scope for creativity on the part of a firm's managers. Given knowledge of the demand and the cost of production, there is only one profit-maximizing price. Once that price is determined (by the simple rule of charging the price that motivates consumers to purchase the quantity where marginal revenue equals marginal cost), the only sensible thing for a manager to do is to charge it. Absolutely no creativity is involved.

In the real world, there is plenty of scope for creative pricing. And such creativity can be very profitable. We have discussed throughout this book how firms compete on many margins. It is common to think of firms competing by producing better products and charging lower prices. And certainly the long-run consequence of firms struggling against each other for more consumer dollars is better products at lower prices. But in this chapter we concentrate on how managers can increase the competitiveness of their firms by producing more creative pricing strategies. Managers can often do as much or more for their firms, and their careers, by coming up with better pricing approaches as by coming up with better products. Of course, as is true of everything else in business, managers must have the proper incentives to be creative in their pricing strategies.

Fair Prices

Real world managers are not limited to charging only one price for a product (although fair trade laws and the penalties that go with their violation do restrict the range of pricing options available to many managers). As those business people who fly frequently know, there are several different prices being charged for a coach seat (or a first-class seat) on most flights. For example, passengers who book their flights weeks in advance often pay less (often several hundred dollars less) than passengers who book their flights days before their departure. By charging different prices for the same product, firms are able to earn higher profits than are possible with only one price. Some creativity can be exercised by carefully announcing prices.

There is a joke based on the pricing creativity of optometrists. When a customer inquires about the price of a pair of glasses the optometrist answers, "seventy five dollars," and then pays close attention to the customer's expression. If he doesn't cringe, the

optometrist quickly adds, "for the lenses." If the customer still doesn't cringe, the optometrist adds, "for each one."[23]

There are, of course, better (perhaps less devious) ways of charging different prices. Consider the demand you face for a book you have written. Let us suppose that you can sell 10,000 copies by charging a price of $25 to everyone, thereby realizing total revenues of $250,000. But you could also sell fewer copies at higher prices. The one person most anxious to read your book is willing to pay $50 for it; the next most anxious reader is willing to pay slightly less for it, on down to the 10,000th most anxious reader who is willing to pay the $25 price at which you could sell all 10,000. You can even sell more copies, of course, but only by charging less than $25.

How should you price your book? If you could somehow charge each reader the maximum amount he or she is willing to pay, then you obviously would not sell each copy for $25. You would charge prices that are higher than $25 on the first 9,999 copies, which would necessarily yield far more in revenues than you would get by charging $25 for each and every copy.[24]

Rarely, if ever, can a seller expect to be able to practice such "perfect" price discrimination (and sellers need to consult their lawyers to insure that they do not violate laws that prevent charging different prices to different customers within the same identified classes of customers that cannot be justified on cost considerations). Even if the demand is known exactly, so the seller knows the maximum amount that can be charged for each unit of the product, the seller is unlikely to be able to identify the consumer who is willing to pay the maximum for each. And if such detailed information were known, it could still be difficult for the seller to charge each consumer a different price because of resale possibilities. For example, if resale is easy (meaning cheap), those who are being charged less than $30 for the book could buy extra copies and sell them to the most anxious readers (who would otherwise be charged more than $45) for $35. Such arbitraging reduces the ability of sellers to profit from price discrimination. But it does not prevent creative managers from finding less than perfect, but still profitable, ways of charging different prices for a product.

Let's return to the book example. Book publishers cannot differentiate between every potential buyer of a book and charge each a different price. But they can separate the market into two broad categories of buyers, those who are most impatient to read the latest novel by, say, Tom Clancy, and those who want to read it but do not mind waiting awhile. If publishers can separate (or segment) these groups, they can charge a different price to each group. But how can they do that? One method: sell hardback and paperback editions of the same book. Hardback books are issued first and are sold at a far higher price than the paperback edition that will not be made available until six

---

[23] As reported by David Friedman, Hidden Order: The Economics of Everyday Life (New York: Harper Business, 1996), p. 134.

[24] By simply assuming that only 10,000 copies of the book are sold, we are ignoring the fact that the number of copies that maximize profits will generally increase when different prices are being charged for the same product. Here we are interested only in pointing out that such price discrimination increases the total revenue received for any given level of sales.

months or a year later. In this way the seller charges those customers who are less sensitive to price (or who have an *inelastic* demand) a higher price than those who are sensitive to price (or who have an *elastic* demand). Those customers with inelastic demands reveal their impatience by their willingness to pay the high hardback price. There is no problem with arbitrage in this case since those who pay the low price do so long after the high-price customers have made their purchases.

Of course, sellers don't always have to package their products differently, as publishers do, to distinguish between buyers who have inelastic demands and those who have elastic demands. Just after new, more powerful models of computers are introduced, their prices can be quite high, only to fall later. Many chalk up the falling prices on computers to reductions in production costs, which may very well be true. However, we suggest an additional explanation for why computer prices fall with the age of the models: the sellers are using time to segment their markets, charging those who are eager to get the new models a high price and charging those who are less eager and more willing to wait a lower price.

Department stores almost always have storewide sales after Christmas. Commonly, the after-Christmas sales are explained by the stores' wanting to get rid of excess inventories. There is a measure of truth to that explanation; stores cannot always judge correctly what will sell in December. However, it is also clear that shoppers have more inelastic demands before Christmas than they have after Christmas. Hence, the stores are often doing nothing more than segmenting their markets. They plan to hold after-Christmas sales and order accordingly. They are not making less money by the sales; they are, in truth, making more money because they can charge different prices in the two time periods, attracting customers they otherwise would have lost.

Grocery stores and the suppliers of the products grocery stores sell have also found a way of getting customers to reveal how sensitive they are to price, which allows those who are less price sensitive to be charged more than those who are more price sensitive. In almost every daily newspaper you can find pages of coupons that, if you cut them out and take them to the designated store, allow you to save 25 cents, 50 cents, and sometimes a dollar or more, on a host of different products. No coupons, no savings.

Those who go to the trouble of cutting out these coupons and carrying them to the store are revealing themselves as being relatively price sensitive. So when you fail to present coupons as you go through the checkout line at your local supermarket, you are telling the cashier that you are not very sensitive to price, that your demand is relatively inelastic. The cashier responds by charging you more for the same products than he or she charged the coupon-laden customer ahead of you. The problem of arbitrage is handled by limiting the amount a customer can buy of a product. Moreover, not many people are tempted by the opportunity to buy one bottle of shampoo for 50 cents off and then trying to sell it for 25 cents off to someone in the parking lot without a shampoo coupon. The cost of creating the secondary market for something as cheap as shampoo is surely greater than the price differential, especially when few units can be bought at the favorable price and sold at a higher price.

Sometimes a firm can profit by charging different prices to different customers without appearing to do so. This can be accomplished by putting the same price on two products that are consumed together by some customers, but not by others. Consider the owner of a theater who realizes that some customers are willing to pay more to go to the movies than others are. Obviously, the owner would like to charge these customers more. But the owner has no way of determining who the price-insensitive customers are when they are paying for their tickets. So how does the manager charge the price-insensitive customers more without losing the remaining customers?

There is a way that we have all observed, but probably didn't think of as an example of price discrimination. Assume that the theater owner believes that those customers who are willing to pay the most to watch a movie are generally the ones who most enjoy snacking while watching. If this assumption is correct (and we will argue in a moment that it probably is), the owner takes advantage of the demand of the enthusiastic movie watchers by charging a moderate price for the tickets to the movie and high prices for the snacks sold in the theater lobby. By keeping the ticket prices moderate the customers with a high demand elasticity for the movie will still buy a ticket since they are not going to do much snacking anyway. While the low elasticity demanders will surely complain about the high prices on all the snacks they eat, they still consider the total cost of their movie experience acceptable since they were willing to pay more for their ticket than they were charged.

If it were not true that those who are willing to pay the most to watch a movie also enjoy snacking the most, then it is unlikely that we would observe such high prices for snacks at the movies.[25] For example, assume that the opposite were true, that those who are not willing to pay much to watch a movie are the ones who enjoy snacking the most when watching the movie. If this were the case, the owner of the theater would find that charging moderate prices for the tickets and high prices for the snacks was not a very profitable strategy. Since the avid movie watchers are not snacking much, they would be willing to pay more than the moderate price to get into the theater. And since the other customers care more about snacking than seeing the movie, they will see little advantage in paying the moderate price for the movie when the snacks are so expensive. In this case, the most profitable pricing strategy would be high-ticket prices and low snack prices. The enthusiastic movie watchers would still come, and end up paying more. And the snackers would now be willing to pay the high-ticket prices for the opportunity to eat lots of cheap snacks.[26] The fact that we do not see such pricing in theaters suggests that, at least for more consumers than not, our assumption is correct.

---

[25] It should be noted that some economists have argued that the high price for snacks at the movie theaters reflect the higher cost of supplying them in movie theaters than in food stores. As opposed to food stores, the snack shop in a movie theater is only open for a limited amount of time during the day. So, as the argument goes, the overhead cost is spread over less time and fewer sales. For an elaboration of this argument, see John R. Lott, Jr. and Russell D. Roberts, "A Guide to the Pitfalls of Identifying Price Discrimination," Economic Inquiry vol. 29, no. 1 (January 1991), pp. 14-23. We do not quarrel with this reasoning, but we also believe that creative price discrimination provides at least part of the explanation for the high price of movie snacks.

[26] Determining the exact combination of prices that maximize profits depends on the relative differences in demand for the two types of customers. If, for example, the avid movie fans were willing to pay a tremendously high price to see the movie and snackers could care less about the movie, but went into frenzies of delight at the mere thought of a Snickers bar, then the best pricing policy would be an extremely high ticket

Any time a firm can identify consumers on the basis of their sensitivity to price, it is in a position to vary its price for different groups in ways that increase the incentive for consumers to purchase its product. The advantage of being able to separate customers willing to pay high prices (again, who have relatively inelastic demands) from those who are more price sensitive (have relatively elastic demands) is so great in some cases that it explains why some firms will incur costs to reduce the quality of their products so they can sell them for less.

For example, soon after Intel introduced the 486 microprocessor it renamed it the 486DX and introduced a modified version, which it named the 486SX. The modification was done by disabling the internal math coprocessor in the original 486, a modification that was costly and reduced the performance of the 486SX. Intel then, in 1991, sold the 486SX for less, $333 as compared to $588 for the 486DX. Why would Intel spend money to damage a microprocessor and then sell it for less?[27] The answer is to separate out those customers who are willing to pay at lot for a microprocessor from those whose demand is more sensitive to price. Intel could sell the 486DX to the former at a price that would have driven the latter to competitive firms. Yet it managed to keep the business of the latter customers by lowering the price to them without worrying that this would drive the price down for the high-end customers. There was no way for the lower-price consumers to buy the lower-price product and sell it to the high-end consumers since its performance had been reduced.

Similarly, when IBM introduced its LaserPrinter E in May 1990, it set the price lower than the price for its earlier model, the LaserPrinter. The LaserPrinter E was almost exactly the same as the LaserPrinter except that the former printed at a rate of 5 pages per minute while the latter printed at a rate of 10 pages per minute. Why was the LaserPrinter E slower? Because IBM went to the expense of adding chips that had no purpose other than to cause the printer to pause so it printed slower. Why did IBM go to extra expense to produce a lower performance printer? Again, to separate its market between consumers with inelastic demands from those with elastic demands so more could be charged for the former than the latter.

One of the authors, Lee, enjoys playing golf (although why he does is a mystery). He buys brand-name golf balls that have been labeled with XXX to indicate that they have some flaw and are sold at a discount. Many good golfers are willing to pay the extra money for regular brand-name balls, which supposedly travel farther than the XXX balls. Lee, on the other hand, sees no advantage in hitting his balls farther into the woods. And anyway, he is not convinced that there really is any difference between the regular high-priced balls and the XXX balls, except the manufacturer went to the extra expense of adding the XXXs. While we have no documentation, we suspect that golf manufacturers simply put XXXs on a certain percentage of their balls so they can

---

price with extremely low-priced (maybe free) snacks. In this case the theater owner would probably stipulate that snack customers would have to eat the snacks in the theater to prevent them from filling large sacks with popcorn and candy bars. This would be no different than the policy of all-you-can-eat restaurants.

[27] It was cheaper to make the 486DX and then reduce its quality than it was to produce the lower quality 486SX directly. This example, the following example, and several other examples of firms intentionally reducing the quality of their products are found in Raymond J. Deneckere and R. Preston McAfee, "Damaged Goods," Journal of Economics and Management Strategy, vol. 5, no. 2 (Summer 1996), pp. 149-174.

separate their market between golfers like Lee, who are quite sensitive to price, and golfers who because they have a reasonable idea where their balls are going, are not very sensitive to price.

Another technique firms can use to separate price-sensitive consumers from those who are less sensitive is to make unadvertised price discounts available, but only to those who search them out and ask for them. Obviously those who go to the trouble to find out about a discount, and then ask for it, are more concerned over price than those who do not. This approach to identifying customers for discounts on long distant calls is (at this writing) being used by AT&T. According to an article in the Wall Street Journal, AT&T responded to Sprint Corporation's 10 cents a minute for calls during weekends and evening hours by offering a flat rate of 15 cents anytime, a plan they called One Rate.[28] But AT&T really had two rates, one of which they did not advertise. The unadvertised rate, available only to those who asked for it, allowed AT&T customers to call around the clock for 10 cents a minute. As reported in the Journal, "AT&T customers can get dime-a-minute calling 24 hours a day, seven days a week -- if they know to ask for it. That is the hardest part, for AT&T has been uncharacteristically quiet about the new offer. The company hasn't advertised the 10-cent rate; it hasn't sent out press releases heralding the latest effort to one-up the folks at Sprint."[29] The old adage about oiling only what squeaks certainly applies in this case. (We suspect that AT&T was not all that pleased with the Wall Street Journal simply because the publicity reduced AT&T's ability to segment its market by reducing the "search costs" that would otherwise have faced AT&T customers who read the Wall Street Journal.)

Sometimes it is possible to charge the same customer more than one price for different units of a product and, by doing so, get the customer to pay more than otherwise. This pricing strategy often works to the firm's advantage even though it is impossible to separate consumers into different groups and charge each a different price, as in the previous examples. The simplest case, both to put into practice and to analyze, involves charging two prices for the same good. Assume that you are selling AA batteries and the cost of producing each of these batteries is 20 cents. Let us suppose that the best you can do when charging one price is to set the price at 60 cents, which allows you to sell 24,000 units. This pricing policy yields a profit per unit of 40 cents (60 cents – 20 cents), which yields a total profit of $9,600 (40 cents x 24,000). But you can raise your profit above $9,600 if, once your customers buy 24,000 batteries at 60 cents each, you can lower the price on any additional batteries they buy. For example, if you reduce the price to 50 cents on all batteries purchased beyond 24,000, you can increase battery sales to, say, 36,000 and make an extra profit on the additional 12,000 units of 30 cents

[28] John J. Keller, "Best Phone Discounts Go to Hardest Bargainers," Wall Street Journal, February 13, 1997, pp. B-1 and B-12.

[29] Ibid., p. B1.

each (50 cents minus 20 cents), which means that profits can go up by $3,600 (30 cents x 12,000).[30]

In fact, firms do use such two-part pricing strategies, but, of course, not exactly in the way just described. For example, if a firm announced that it was going to charge 60 cents for each battery until the first 24,000 were sold each week, and then charge 50 cents per battery for the remainder of the week, it probably would not sell the 24,000 since everyone would attempt to postpone their purchases until enough other consumers had made theirs. But a firm can effectively achieve much the same result by making everyone the following offer: buy two batteries at a price of 60 cents and get the third for 50 cents. Such a two-part pricing offer is easy to implement and can increase profits. Not surprisingly, such offers are commonly observed.

The more competition, and price rivalry, in an industry the smaller the gain a firm in that industry can realize from charging different customers different prices. Even relatively price-insensitive customers will be bid away by rival firms when price competition is intense, if one firm tries to charge those customers much more than it does its more price sensitive customers. Nevertheless, the more the firms in an industry can segment their market so as to buffer the price competition between them, the greater the scope for creative pricing strategies that can increase profits, a point to which we can now turn.

Cartel Cheating

Firms in an industry could simply get together and agree not to compete consumers away from each other by reducing prices. This would allow them to keep prices, and their collective profits, higher than would be possible if all firms made a futile attempt to increase their market shares by charging lower prices. But there are two problems with this approach to reduce price competition. The first problem is that any agreement to restrict competition *can be* illegal, and firms and their managers, who enter into such an agreement, risk harsh antitrust penalties. The second problem is that even if agreements to restrict price competition were not illegal, they would still be almost impossible to maintain. Members of industry cartels that have agreed to set prices above competitive levels are in another prisoners' dilemma. While they are collectively better off when everyone abides by the agreement, each individual sees the advantage in reducing price below the agreed upon amount. If other firms maintain the high price, then the firm that cheats on the agreement can capture lots of additional business with a relatively small decrease in its price. On the other hand, if the other firms are expected to cheat on the agreement, it would be foolish for a firm to continue with the high price since that firm would find most of its customers competed away. Only if firms ignore prisoners-dilemma temptations, and take the risk of making the cooperative choice, can cartel price agreements be maintained. Not surprisingly, such agreements tend to break down.

---

[30] As opposed to what many may think, the higher profits from creative pricing do not necessarily come at the expense of consumers. In the situation just described, consumers are also better off to the extent that they value each of the additional 12,000 batteries more than the price they pay for them.

The Organization of Petroleum Exporting Countries (OPEC) is a classic example of a cartel with all the hopes and dreams of a well-oiled cartel but with rampant cheating. What is amazing is that the cartel held together for as long as it did in the 1970s. Now, it pretends to set production restrictions only to have them violated fragrantly. One unheralded explanation for Sadam Husein's invasion of Kuwait was that prior to its invasion of Kuwait, Iraq had been trying to hold to its assigned quota while Kuwait flagrantly violated its quota, denying Iraq sales than the higher world oil price the cartel sought. By taking over Kuwait and (possibly) then Saudi Arabia, Hussein could introduce some needed production discipline into the cartel and raise the world price of oil, a threat that helps to explain why the industrial countries, including the United States, were willing to militarily defend Kuwait. The allied forces were, in effect, trying to maintain the natural competitive instability in the cartel (as well as trying to deny a tyrant greater political clout on the world stage).

Pricing Strategies that Moderate Price Competition

There are pricing policies, however, that can moderate price competition between rival firms without the need for a cooperative agreement. Ironically, these strategies do more to reduce competition when competition motivates most firms in an industry to implement them once the managers in one firms does.

Consider a pricing policy that would seem to favor your customers with protection against high prices but which is a smart policy because it makes higher prices possible. The strategy is quite simple, involving an unqualified pledge, "We will meet or beat any competitor's price." A so-called "meet-the-competition" pricing policy tells your customers that if a competitor offers them a lower price, you will match it. This policy is commonly advertised as "guaranteed lowest prices," by retail stores like Circuit City and many others. To implement such a policy you inform your customers that if they can find a lower price on a product within thirty days of purchasing it from you, they will receive a rebate equal to the difference. Obviously such price guarantees have value to the customers, but what is not widely appreciated is that the guarantees, especially if they are also made by those you are in competition with, allow you to charge more than otherwise. How can this be?

One straightforward explanation is that the price assurance gives customers some insurance and, because of that added attribute, increases their demand. The greater demand leads to higher prices.

But there is another explanation based on an equally simple proposition: if you want to charge higher prices there is an obvious advantage in discouraging competitors from reducing their prices to compete your customers away. This is exactly what a meet-the-competition policy does. Your competitors are probably not all that anxious, in any event, to initiate a price-cutting campaign. Attempting to compete customers away from another firm through lower prices is always costly. If successful, the new business is likely to be worth less to the pricing-cutting firm than to the firm that loses it because the

price is now lower. Also, existing customers will want to receive a lower price as well, which can cut deeper into any profits that might have otherwise been possible. Of course, if a price-cutting campaign aimed at capturing new customers fails to do so, the campaign is all cost and no benefit. So if your competitors know that you have a meet-the-competition agreement with your customers they will have less, and likely nothing, to gain from trying to attract those customers by cutting their prices.

A meet-the-competition pricing policy cannot only be good for your profits, it can also be good for your competitors as well. By allowing you to keep your prices higher than otherwise, your meet-the-competition policy gives your competitors more room to keep their prices high. This suggests that, as opposed to most competitive strategies that become less effective when mimicked by the competition, your meet-the-competition policy becomes more profitable when other firms in the industry implement the same policy. Just as your competitors are better off when you do not have to worry about the competitive consequences of keeping your prices high, so are you better off when your competitors are relieved of the same worry.[31]

A related pricing policy is to offer some of your customers the status of most-favored-customer, which entitles them to the best price offered anyone else. (Again, this policy must be checked with lawyers, given that some such policies in some circumstances might be construed as illegal.) If you lower your price to any customer, under this policy you are obligated to lower it for all of your most-favored customers. As with the meet-the-competition policy, what at first glance appears to favor your customers can actually give the advantage to you. A most-favored customer policy increases the cost of trying to compete customers away from rival firms by reducing price. And when one firm has such a policy, its reluctance to engage in price competition makes it easy for other firms to keep their prices high. So, as with meet-the-competition policy, the advantage firms realize from a most-favored-customer policy is greater when all the firms in an industry have such a policy.

If the idea that a policy of being quick to reduce prices for your customers can result in higher prices seems counter-productive, you are in good company. In their book on Co-opetition, Harvard business Professor Adam Brandenburger and Yale management Professor Barry Nalebuff relate how Congress, in an effort to control the cost of campaigning, required television broadcasters to make candidates for Congress most favored customers. In the 1971 Federal Election Campaign Act, Congress made it against the law for TV broadcasters to lower their rates for a TV spot to any commercial customer without also lowering their rates to candidates. The result was that TV broadcasters found it extremely costly to reduce rates for anyone, and the networks made more money than ever before. Politicians had the satisfaction of knowing that they did not pay more for airtime than anyone else, but they likely ended up paying more (as commercial advertisers surely did) than they would have without forcing the broadcasters to implement a most-favored-customer pricing policy.

---

[31] Our discussion of meet-the-competition pricing is based on Chapter 6 of Barry J. Nalebuff and Adam M. Brandenburger, Co-opetition (London: HarperCollins Business, 1996). Our subsequent discussion of most-favored-customer policies and preferred customer discounts also draw heavily from Nalebuff and Brandenburger's excellent book.

Congress made a similar mistake in 1990 when it attempted to reduce government reimbursements for drugs by stipulating that Medicaid would pay only 88 percent of the average wholesale price for branded drugs, or, if lower, the lowest price granted anyone in the retail trade drug business. But instead of lowering prices, the law actually raised them. By making itself a most-favored customer, the federal government gave the drug companies a strong incentive to raise prices for everyone. And indeed that is exactly what happened, according to a study cited by Nalebuff and Brandenburger that found that prices on branded drugs increased from 5 to 9 percent because of the 1990 rule changes.[32] The advantage the government may have realized by keeping its price down to 88 percent of the average wholesale price was probably more than offset (it was often receiving a discount anyway) by the higher average prices. And certainly non-Medicare patients ended up paying higher drug prices, a disguised form of what NBC News should surely want to cover under its "Fleecing of America" segment.

Advantages of Frequent Flyer Programs

Another pricing strategy that allows the firms in an industry to reduce price competition has become increasingly common in recent years. This strategy involves a creative way of identifying those customers who are most likely to buy from your firm anyway and then lowering the price they pay. At first glance such a strategy would appear counterproductive. Why would you lower the price for those who are likely to buy from you even if you charge a higher price? The answer is that by making price concessions to your most loyal customers you can end up charging them higher prices.

A good way of explaining this seemingly paradoxical possibility is by considering the frequent-flyer programs that almost all the airlines now have. These programs are commonly thought of as motivated by each airline's desire to compete business away from other airlines by effectively lowering ticket prices. No doubt this was the primary motivation when, in 1981, American Airlines introduced its AAdvantage program. And the rapidity with which other airlines countered with their own frequent-flyer programs suggests intense competition between the airlines. But intended or not, the proliferation of these programs has had the effect of reducing the direct price competition between airlines and, as a result, may be allowing them to maintain higher prices than would otherwise be possible. An airline's frequent-flyer program reduces the effective, if not the explicit, price it charges its most loyal customers, and reinforces their loyalty.[33] By increasing the motivation of an airline's frequent flyers to concentrate their flying on that airline, it decreases the payoff other airlines can expect from trying to compete those customers away with fare reductions. This allows the airline with the frequent-flyer program to keep its explicit fares higher than if other airlines were aggressively reducing

---

[32] Ibid., pp. 164-165.

[33] Even when a person is a member of more than one frequent-flyer program, there is an advantage in concentrating patronage on one airline since the programs are designed to increase benefits more than proportionally with accumulated mileage.

theirs.[34] This decreased motivation to engage in price competition becomes mutually reinforcing as more airlines implement frequent-flyer programs.

From the perspective of each airline it would be nice to be able to compete away customers from other airlines with lower fares, but collectively the airlines are better off by reducing this ability. And this is exactly what the spread of frequent-flyer programs has done, to some degree, by segmenting the airline market. There is now less competitive advantage in reducing airfares, and less competitive disadvantage in raising them. The effect has been to reduce the elasticity of demand facing each airline, which allows all airlines to charge higher prices than would otherwise be sustainable.[35]

A pricing strategy similar to frequent-flyer programs has begun to spread in the automobile industry. In 1992 General Motors joined with MasterCard and issued the GM credit card. By using the GM card a consumer earns a credit equal to 5 percent of their charges that can be applied to the purchase or lease of any new GM vehicle (with a limit of $500 per year up to $3,500 for any one purchase). While not all the major automakers have followed the GM lead, several have. And the more automakers that join in, the better for the car industry in general. Just like frequent-flyer programs, automobile credit cards allow a car company to focus implicit price reductions on its most loyal customers. An individual is not likely to be using a GM credit card unless she is planning on buying a GM car or truck. As the number of car companies that issue their own credit card increases, the more the auto market will become segmented and the less the advantage from price competition. Again, a pricing policy that allows a firm to target its more loyal customers and favors them with price cuts can have the effect of increasing the prices being charged.

Saying that firms *should* come up with creative pricing schemes is easier said than done. Managers must have the right incentives to do it. If an organization only offers rewards for developing new product lines or for getting workers to increase production of the given product lines, managers may overlook equally effective alternative ways to increase profits. Firms would be well advised to use profit as a prominent performance measure simply because it gives managers flexibility to look for profits in all kinds of ways, in the way products are developed and marketed and in the way they are priced.

* * * * *

---

[34] You may be thinking that keeping the explicit fares higher does not mean much if, because of the frequent-flyer programs, the actual fares to customers are lower because of the value of their mileage awards. But one of the big advantages of frequent-flyer programs is that they do not cost the airlines as much as they benefit the customer. Flights are seldom completely sold out, so most of the free flights awarded end up filling seats that are unsold. Of course, frequent flyers do use their mileage for flights they would have otherwise paid for. But by allowing frequent flyers to transfer their mileage awards to others, say a spouse or child, the airlines increase the probability that those who would not have otherwise bought a ticket will use those awards.

[35] Another way of seeing the advantage of segmenting the market is by recognizing that reducing the elasticity of demand facing each airline also reduces the marginal revenue of each airline and brings it more in line with the marginal revenue for the industry. The closer each firm's marginal revenue is to the industry's marginal revenue, the closer the independent pricing decisions of each firm in the industry will come to maximizing their collective profits.

We cannot exhaust the possibilities for creative pricing policies in this "Manager's Corner." We have, however, indicated some of the ways that managers can increase the profitability of their firms by taking full advantage of the subtle interactions between incentives and pricing policies.

Lower prices surely increase the incentive a consumer has to buy your product. However, some customers have a stronger incentive to take price into consideration than others do, and these different price sensitivities create profitable opportunities to charge different prices for the same product. Such opportunities are greater the less the danger of your customers being captured by the aggressive price cutting of rival firms. Fortunately there are pricing policies that can reduce that danger. Such policies as meet-the-competition pricing can reduce the incentive other firms have to engage in price competition. Other policies, such as those represented by frequent-flyer programs, reduce price competition by reducing customer incentives to take the price (at least the explicit price) into account. By tailoring such pricing strategies to their particular circumstances, managers can do what good managers are paid to do: use incentives to increase the profitability of their firms.

Of course, managers should be given an incentive to consider the profitability of devoting attention to pricing as well as to other ways of increasing the profitability of their firms. We suspect that the American Airlines manager who came up with the idea of the AAdvantage frequent-flyer program has been handsomely rewarded for his or her creativity. When a pricing innovation is as distinctive and profitable as the AAdvantage program has been, it is easy to recognize and reward those who are responsible. But few pricing innovations will have the bottom-line impact that the AAdvantage program had for American Airlines, with it more difficult to sort out how important a particular contribution is. Rewarding managers for more creative pricing strategies is best done in the same way they are rewarded for all the many marginal things they do to improve their firm's profitability -- tie their compensation to that profitability. The closer managerial compensation comes to creating the incentives of a residual claimant, the more alert managers will be to adding value along the entire spectrum of possibilities, from coming up with better products, developing less costly ways of producing those products, and devising more creative ways to price them.

MANAGER'S CORNER II: **The Desktop Monopoly**

In its antitrust case against Microsoft, the Justice Department has charged that Microsoft's monopoly is nowhere more evident than in its control of the "desktop," or the first screen in view after Windows has booted. By its control of the desktop, the Justice Department contends, Microsoft has been able to spread the use of its own web browser, Internet Explorer, while curbing the use of competitor Netscape's browser, Navigator.

How has Microsoft done this? By not placing an icon for Netscape's Navigator on the desktop. The Justice Department reasons that Microsoft should, in any settlement of the current antitrust suit, be forced to place an icon for Navigator on the desktop.

Is this reasonable? First, it must be understood that Microsoft does not prevent Netscape from having its icon on the desktop. There are two possible ways in which

Netscape can get its icon there. First, Microsoft does not deny computer manufacturers the right to put additional icons on the desktop before they ship their computers to customers. Netscape can have computer manufacturers install its icon. All Netscape has to do is pay the requisite price for their doing that. Second, Netscape can pay Microsoft to put its icon on the desktop on all versions of Windows that Microsoft ships to retailers and to computer manufacturers, which will then install that version on the computers they ship. Microsoft has indicated a willingness to make such deals. It has placed AOL's icon on the desktop – in exchange for AOL's agreement to make Internet Explorer its recommended browser. If Netscape doesn't make a similar deal with Microsoft or computer manufacturers, then that appears to be reason enough to suspect that Netscape just doesn't want to pay up for what is reputed to be the "most valuable real estate" in the world. We come to Microsoft's defense because we don't believe that Netscape should be allowed to use the powers of the Justice Department to get something for nothing.

There is, in short, a considerable measure of unfairness in the Justice Department's proposal. We think Phil Lemmons, editorial director of PC World and an advocate for more choice in operating systems (which means not always a friend of Microsoft), made an important point too easily forgotten in the rush to get at Microsoft: "In essence, they [Microsoft's critics and Justice Department lawyers] want to compel Microsoft to distribute, within Windows itself, products that will compete with Windows. This stealthy approach achieves the noteworthy feat of treating a monopoly unfairly. It's as though AMD, Cyrix, and IDT demanded that Intel embed their instructions in the Pentium II."[36]

In addition, if the Justice Department forces Microsoft to install a Netscape icon on the desktop, such a condition of settlement would solve nothing for very long, a point that Justice Department lawyers, who must know the work of Ronald Coase, should appreciate. Coase reasoned decades ago that in the absence of high costs of negotiating the exchange of property, the ultimate distribution of property would be little affected by how the property is initially distributed.[37] Those who put the highest value on it would ultimately hold any given piece of property. This has come to be widely known as the Coase Theorem.

For example, suppose that Sam owns a given acre of land that he values at $100,000. Suppose also that Sue could use the land more profitably and, hence, values it at $150,000. What would happen? Sam would sell the land to Sue at a price between $100,000 and $150,000, and both Sam and Sue would be better off. If Sue owned the property initially, the property would remain with her. Sam could not cover Sue's cost of $150,000.

This very simple line of argument is fully applicable to the desktop and the Microsoft case. Clearly, Microsoft should be willing to sell space on the desktop at some price, as it did with AOL. If Netscape does not have an icon on the Windows desktop, it

---

[36] Phil Lemmons, "Flattery Will Get You Bad Publicity," PC World, June 1998, p. 19.

[37] Ronald H. Coase (1964), "The Problem of Social Cost," Journal of Law and Economics, vol. 3 (October), pp.1-44.

must be because Microsoft has placed a value on not having the Netscape icon on the desktop that is higher than Netscape's value of having it there. If the reverse were true (Netscape valued the desktop space more highly than Microsoft), then Netscape would have bought a place on the desktop for its icon long ago.

If the Justice Department gets its way and Microsoft is forced to have a Navigator icon on the desktop, Coase's central point still holds. The value Microsoft places on not having the Navigator icon on the desktop should still be higher than the value Netscape places on having the icon there. As a consequence, once the dust from the trial settles, Microsoft should quickly pay Netscape to remove its icon. The only meaningful lasting change would be that some of Microsoft's wealth is transferred to Netscape.

The Justice Department may reason that this outcome represents "justice," given Microsoft's alleged monopoly power to dictate market outcomes, including use of the desktop. While the monopoly claim is surely disputable, all that needs to be pointed out here is that the ruling does nothing to thwart Microsoft's supposed monopoly powers. It means, however, that Netscape, which invested nothing to develop the Windows operating system network and desktop, will have gotten Microsoft's property for nothing. It also means that Microsoft will be forced to use whatever market power, as well as its expertise in developing and promoting programs, to buy back its property, in the process possibly hiking its prices (albeit marginally) in order to pad the pockets of Netscape's owners. It is hard for us to see how such an outcome constitutes "justice" or protects consumers.

Concluding Comments

The consequences of monopoly are higher prices and lower production levels than are possible under perfect competition. Monopoly power can also result in inefficiency in production, for the monopolistic firm does not produce to the point where its marginal cost equals the consumer's marginal benefit. Consumers might prefer that more resources be used in the production of a monopolized good and might be willing to pay a price that exceeds the cost of production for additional units of the good. However, the profit-maximizing monopolist stops short of that point.

The new "network economy" often times turns much economic analysis on its head. This is especially true when it comes to discussions of "monopoly power." A market for a network good might tend toward a single seller. At the same time, that single seller may have no, or very little, control over market price, mainly because of the network effect. And a firm producing a network good can easily justify selling the good at zero (or below-zero) prices.

The next chapter examines the two remaining market structures, monopolistic competition and oligopoly. Much of the analysis in this chapter is applicable to those market structures, for each has a degree of monopoly power. The power of the monopolistic competitor and the oligopolist is circumscribed by the existence of other firms in the industry, however, and by the fact that other firms may enter the market.

Review Questions

1. Many magazines offer multiyear subscriptions at a lower rate than one-year subscriptions. Explain the logic of such a scheme. Why might it be considered evidence of monopoly power on the part of the magazines?

2. Explain why a monopolized industry will tend to produce less than a competitive industry.

3. "If a monopoly retains its market power over the long long-run, it must be protected by barriers to entry." Explain. List some restrictions on the mobility of resources that might help a firm retain monopoly power.

4. Why, from an economic point of view, should antitrust action not be taken against all monopolies?

5. Given the information in the table below, complete the monopolist's marginal cost and marginal revenue schedules. Graph the demand, marginal cost, and marginal revenue curves, and find the profit-maximizing point of production. Assuming this monopolistic firm faces fixed costs of $10, and must charge the same price for all units sold, how much profit does it make?

Quantity Produced and Sold	Price	Total Variable Cost	Marginal Cost	Marginal Revenue
1	$12	$ 5		
2	11	9		
3	10	14		
4	9	20		
5	8	28		
6	7	38		

6. On the graph developed for question 5, identify the output and profits of a monopolist capable of perfect price discrimination.

7. Suppose a monopoly capable of imperfect price discrimination divides its market into two segments, as shown in graphs (a) and (b). In graph (c), draw the monopolist's combined marginal revenue curve. Then, using the monopolist's marginal cost curve, as shown in graph (c), determine the monopolist's profit-maximizing output level. Indicate the quantity and price of the product sold in each market segment.

8. If a buyers fear that a "network firm" will become a true monopolist in the future, what does that fear do to the firm's current pricing policies?

9. What is the impact of antitrust enforcement in a market for a network good?

Appendix: Marginal Revenue Curve, A Graphical Derivation

Demand curves can be *linear* or *nonlinear*. Once we have learned how to derive the MR curve for the linear demand curve, we can readily adapt the procedure to derive the MR curve for the nonlinear demand curve.

Linear Demand

The graphic derivation of the marginal revenue curve corresponding to a linear demand curve is easy to present. From our examination of marginal revenue in an earlier chapter, we know that

$$MR = P\left(1 - \frac{1}{E}\right)$$

where P is the price and E is the absolute value of the price elasticity of demand. Because the price elasticity of demand is infinite at the point of intersection of the demand curve and the vertical price axis, we know that $1/E = 0$ at the vertical intercept and $MR = P$.

We have now established one point on the MR curve. Since the MR curve for a linear demand curve is also linear,[38] we need to determine only one additional point to construct the MR curve. The second point can be easily determined by setting Equation (9-1) equal to 0 and solving for E, which gives us

$$P\left(1 - \frac{1}{E}\right) = 0$$

$$1 - \frac{1}{E} = 0$$

$$E = 1$$

Thus, when $MR = 0$, $E = 1$. Recall from an earlier chapter that the price elasticity of demand is equal to 1 at the midpoint of a linear demand curve. The point on the horizontal axis corresponding to $E = 1$ on the demand curve will be one-half the distance between the origin and the horizontal intersection of the demand curve. Since MR = 0

---

[38] This result can be shown with the aid of calculus. Given the linear demand curve

$$P = a - bQ$$

Total revenue is

$$TR = PQ = (a - bQ)\,Q = aQ - bQ^2$$

And marginal revenue is

$$\frac{dTR}{dQ} = a - 2bQ$$

Thus, the MR curve is linear, intersects the vertical axis at a (the demand curve's intercept), and has an absolute slope two times that of the demand curve.

when $E = 1$, the second point on the MR curve will lie one-half the distance between the origin and the horizontal intercept of the demand curve.

Our conclusions concerning the shape and the location of the MR curve are illustrated in Figure 12.12. The linear demand curve intersects the vertical price axis at point *P*, and this point is also the vertical intercept of the MR curve. Halfway down the demand curve, $E = 1$ at point *B*, which corresponds to Q_1 on the horizontal axis. Point Q_1, in turn, is midway between the origin and Q_2, which is the horizontal intercept of the demand curve. The *MR* curve is the heavy dashed line connecting point *P* and Q_1 in Figure 12.13.

Since the *MR* curve and the demand curve have the same vertical intercept and the horizontal intercept of the *MR* curve is one-half that of the demand curve, it follows that the slope of the *MR* curve will be two times the slope of the demand curve.[39]

The fact that the slope of the *MR* curve is twice the slope of the demand curve provides us with an alternative method for graphically determining the marginal revenue at any level of output. To illustrate this method, suppose that we wish to determine *MR* at output Q_0, which corresponds to point *C* on the demand curve in Figure 12.13. We accomplish this simply by drawing a horizontal line from point *C* to point *D* on the vertical axis. Bisecting the line *DC* gives us point *F*. A straight line drawn from the vertical intercept through point *F* has exactly twice the slope of the demand curve and is therefore the *MR* curve. The intersection of the *MR* curve with dashed line CQ_0 at point *G* gives us the value of the marginal revenue (read off the horizontal axis) corresponding to point *C*. Although this technique is somewhat laborious, it is useful in graphing the MR curve corresponding to a nonlinear demand curve.

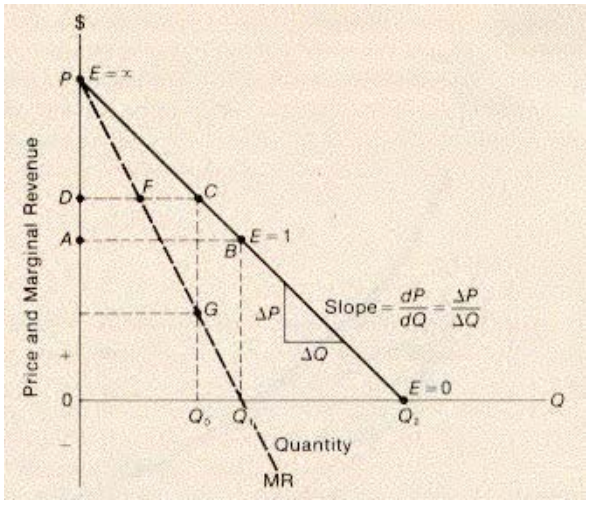


Figure 12.13 Construction of the Linear Marginal Revenue Curve

The marginal revenue curve always starts the intersection of the vertical axis and any demand curve. However, for a linear demand curve, the marginal revenue curve must slope downward under the demand curve, splitting the horizontal distance between the vertical axis and every point on the demand curve. The marginal revenue curve must cut the horizontal axis at the point below the middle of the linear demand curve, or where the elasticity coefficient equals 1.

[39] From the figure, we know that the slope of the demand curve is P/Q_2 and the slope of the MR curve is P/Q_1. Since $Q_1 = 1/2\ Q_2$, the slope of the MR curve is therefore $P/1/2\ Q_2$ or $2P/Q_2$, which is twice the slope of the demand curve. See also footnote above.

Nonlinear Demand

When the demand curve is nonlinear, such as curve *DD* in Figure 12.14, the *MR* curve is constructed using a variation of the technique we have just learned. Essentially, we determine the marginal revenues corresponding to several points on the demand curve and then connect these points with a smooth curve to obtain the MR curve.

A line originating on the vertical axis at point V_1 is drawn tangent to point *A* on the demand curve in Figure 12.14. If we assume that this tangent line is a linear demand curve, then the marginal revenue of this demand curve at point *A* is identical to the marginal revenue of the nonlinear demand curve at point *A*, because the slopes of the two demand curves are equal at point *A* and have the same corresponding price P_1 and quantity Q_1. Therefore, to determine the marginal revenue graphically, we simply draw a straight line from V_1 that bisects line P_1A. This line intersects line AQ_1 at point *B*, giving us the marginal revenue that corresponds to point *A* on the demand curve.

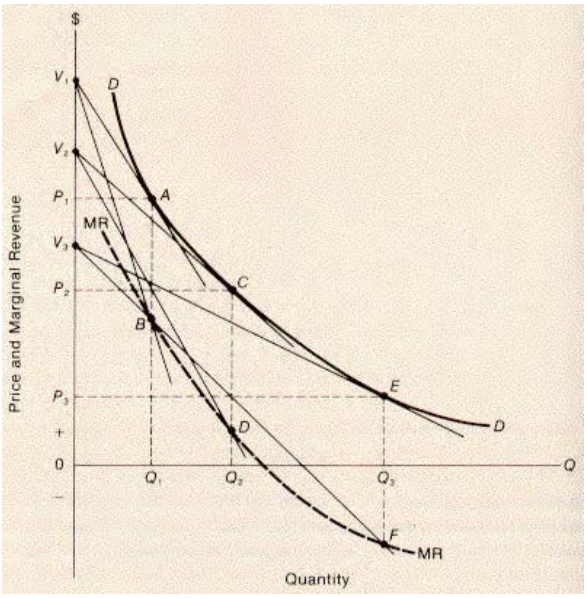


Figure 12.14 Construction of the Nonlinear Marginal Revenue Curve

The marginal revenue curve for a nonlinear demand curve is obtained by imagining linear demand curves tangent to every point on the nonlinear demand curve and finding the midpoint between the vertical axis and the imagined linear demand curves.

Point *B* is the only point on the *MR* curve associated with the nonlinear demand curve *DD*. To construct this *MR* curve, we must determine the marginal revenues that correspond to additional points on curve *DD*. Points *D* and *F* on the *MR* curve are determined for points *C* and *E* on curve *DD* by repeating the steps we followed to locate point *B*. The construction lines required to obtain points *D* and *F* are drawn in the figure, and you should verify that these points have been correctly determined.

Once a sufficient number of points on the *MR* curve have been located, a smooth curve drawn through these points is the graphically constructed *MR* curve associated with the nonlinear demand curve. Figure 12.14 shows that this *MR* curve is also nonlinear and lies below the demand curve.

CHAPTER 13

Imperfect Competition and Firm Strategy

Differences in tastes, desires, incomes and locations of buyers, and differences in the use which they wish to make of commodities all indicate the need for variety and the necessity of substituting for the concept of a "competitive ideal," an ideal involving both monopoly and competition.

Edward Chamberlin

We have so far considered two distinctly different market structures: perfect competition, characterized by producers that cannot influence price at all because of extreme competition; and pure monopoly, in which there is only one producer of a product with no close substitutes and whose market is protected by prohibitively high barriers to entry. Needless to say, most markets are not well described by either of those theoretical structures. Even in the short run, producers typically compete with several or many other producers of similar, if not identical, products. General Motors Corporation competes with Ford Motor Company, Chrysler Corporation, and a large number of foreign producers. McDonald's Corporation competes with Burger King Corporation, Hardees, and a lot of other burger franchises, as well as with Pizza Hut, Popeye's Fried Chicken, and Long John Silver's. People's Drug stores compete directly with other drug chains and locally owned drugstores, and indirectly with department and discount stores that sell the same non-drug products. In the long run, all these firms must compete with new companies that surmount the imperfect barriers to entry into their markets. In short, most companies competing in the imperfect markets can cause producers to be more efficient in their use of resources than under pure monopoly, although less efficient than in perfect competition. One word of caution, however: The study of so-called real-world market structures can be frustrating. Although models may incorporate more or less realistic assumptions about the behavior of real-world firms, the theories developed from them are conjectural. At best, they allow economists to speculate on what may happen under certain conditions. Real-world markets are imperfect, complex phenomena that often do not lend themselves to hard-and-fast conclusions.

Monopolistic Competition

As we have noted in our study of demand, the greater the number and variety of substitutes for a good, the greater the elasticity of demand for that good—that is, the more consumers will respond to a change in price. By definition, a monopolistically competitive market like the fast-food industry produces a number of different products,

most of which can substitute for each other. If Burger Bippy raises its prices, then, consumers can move to another restaurant that offers similar food and service. Because of consumer ignorance and loyalty to the Big Bippy, however, Burger Bippy is unlikely to lose all its customers by raising its prices. It has some monopoly power. Therefore, it can charge slightly more than the ideal competitive price, determined by the intersection of the marginal cost and demand curves. Burger Bippy cannot raise its prices too much, however, without substantially reducing its sales.

The degree to which monopolistically competitive prices can stray from the competitive ideal depends on

- the number of other competitors
- the ease with which competing firms can expand their businesses to accommodate new customers (the cost of expansion)
- the ease with which new firms can enter the market (the cost of entry)
- the ability of firms to differentiate their products, by location or by either real or imagined characteristics (the cost differentiation)
- public awareness of price differences (the cost of gaining information on price differences)

Given even limited competition, the firm should face a relatively elastic demand curve—certainly more elastic than the monopolist's.

Monopolistic Competition in the Short Run

In the short run, a monopolistically competitive firm may deviate little from the price-quantity combination produced under perfect competition. The demand curve for fast-food hamburgers in Figure 13.1 is highly, although not perfectly, elastic. Following the same rule as the perfect competitor and pure monopolist, the monopolistically competitive burger maker produces where $MC = MR$. Because the firm's demand curve slopes downward, its marginal revenue curve slopes downward too, like the pure monopolist's. The firm maximizes profits at M_{mc} and charges P_{mc}, a price only slightly higher than the price that would be achieved under perfect competition (P_c). (Remember, the perfect competitor faces a horizontal, or perfectly elastic, demand curve, which is also its marginal revenue curve. It produces at the intersection of the marginal cost and marginal revenue curves.) The quantity sold with monopolistic competition is also only slightly below the quantity that would be sold under perfect competition, Q_c. Market inefficiency, indicated by the shaded triangular area, is not excessive.

The firm's short-run profits may be slight or substantial, depending on demand for its product and the number of producers in the market. In our example, profit is the area bounded by $ATC_1 P_{mc} ab$, found by subtracting total cost ($0ATC_1 bQ_{mc}$) from total revenues ($0P_{mc} aQ_{mc}$), as with monopolies.

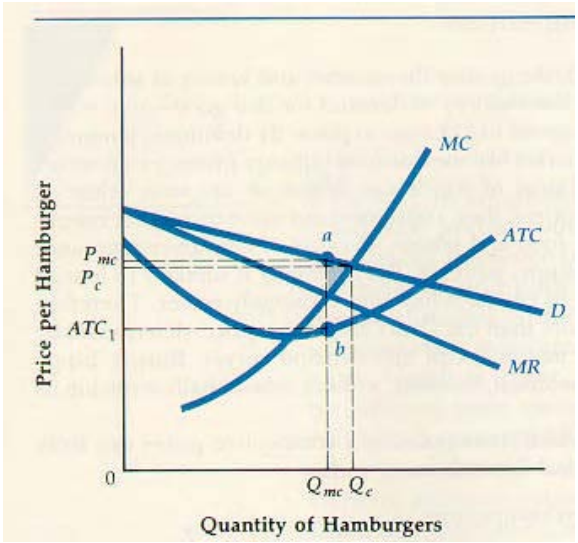


FIGURE 13.1 Monopolistic Competition in the Short Run

Like all profit-maximizing firms, the monopolistic competitor will equate marginal revenue with marginal cost. It will produce Q_{mc} units and charge price P_{mc}, only slightly higher than the price under perfect competition. (The perfect competitor's combined demand and marginal revenue curve would be horizontal at price P_c.) The monopolistic competitor makes a short-run economic profit equal to the area $ATC_1P_{mc}ab$. The inefficiency of its slightly restricted production level is represented by the shaded triangular area.

Monopolistic Competition in the Long Run

Because the barriers to entry into monopolistic competition are not excessively costly to surmount, substantial short-run profits will attract other producers into the market. When the market is divided up among more competitors, the individual firm's demand curve will shift downward, reflecting each competitor's smaller market share. As a result, the marginal revenue curve will shift downward as well. The demand curve will also become more elastic, reflecting the greater number of potential substitutes in the market. (These changes are shown in Figure 13.2.) The results of the increased competition are:

- The quantity produced falls from Q_{mc2} to Q_{mc1}.
- The price falls from P_{mc2} to P_{mc1}.

Profits are eliminated when the price no longer exceeds the firm's average total cost. (As long as economic profit exists, new firms will continue to enter the market. Eventually the price will fall enough to eliminate economic profit.)[1]

Notice that the firm is not producing and pricing its product at the minimum of its average total cost curve, as the perfect competitor would (nor did it in the short run).[2] In this sense the firm is producing below capacity, by $Q_m - Q_{mc2}$ units.

[1] The monopolistic competitor will still have an incentive to stay in business, however. It is economic profit, not book profit, that falls to zero. Book profit will still be large enough to cover the opportunity cost of capital plus the risk cost of doing business.

In terms of price and quantity produced, monopolistic competition can never be as efficient as perfect competition. Perfectly competitive firms obtain their results partly because all producers are producing the same product. Consumers can choose from a great many suppliers, but they have no product options. In a monopolistically competitive market, on the other hand, consumers must buy from a limited number of producers, but they can choose from a variety of slightly different products. For example, the pen market offers consumers a choice between felt-tipped, fountain, and ballpoint pens of many different styles. This variety in goods comes at a price—the higher price illustrated in Figure 13.2.

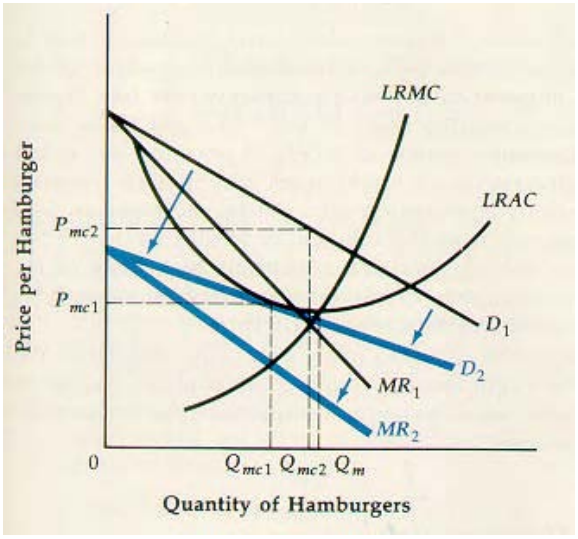


FIGURE 13.2 Monopolistic Competition in the Long Run

In the long run firms seeking profits will enter the monopolistically competitive market, shifting the monopolistic competitor's demand curve down from D_1 to D_2 and making it more elastic. Equilibrium will be achieved when the firm's demand curve becomes tangent to the downward-sloping portion of the firm's long-run average cost curve. At that point, price (shown by the demand curve) no longer exceeds average total cost; the firm is making zero economic profit. Unlike the perfect competitor, this firm is not producing at the minimum of the long-run average total cost curve. In that sense it is underproducing, by $Q_m - Q_{mc2}$ units.

Oligopoly

In a market dominated by a few producers, where entry is difficult—that is, in an oligopoly—the demand curve facing an individual competitor will be less elastic than the monopolistic competitor's demand curve (see Figure 13.3). If General Electric Company raises its price for light bulbs, consumers will have few alternative sources of supply. A price increase is less likely to drive away customers than it would under monopolistic competition, and the price-quantity combination achieved by the company will probably be further removed from the competitive ideal. In Figure 13.3, the oligopolist produces only Q_o units for a relatively high price of P_o, compared with the perfect competitor's price-quantity combination of Q_{cPc}. The shaded area representing inefficiency is fairly large.

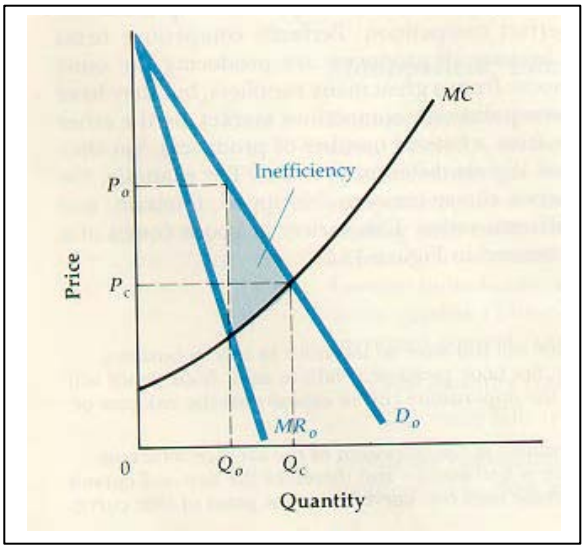
Exactly how the oligopolist chooses a price is not completely clear. We will examine a few of the major theories proposed. In contract, we had to examine only a

[2] The perfect competitor produces at the minimum of the average total cost curve because its demand curve is horizontal—and therefore the demand curve's point of tangency with the average total cost curve is the low point of that curve.

single theory each for perfect competition, pure monopoly, and monopolistic competition.

FIGURE 13.3 The Oligopolist as Monopolist

With fewer competitors than the monopolistic competitor, the oligopolist faces a less elastic demand curve, D_o. Each oligopolist can afford to produce significantly less -- Q_o -- and to charge significantly more than the perfect competitor, who produces Q_c, at a price of P_c. The shaded area representing inefficiency is larger than that of a monopolistic competitor.



Theories of Price Determination

Because each oligopolist is a major factor in the market, oligopolists' pricing decisions are mutually interdependent. The price one producer asks significantly affects the others' sales. Hence when one oligopolistic firm lowers it price, all the others can be expected to lower theirs, to prevent erosion of their market shares. The oligopolist may have to second-guess other producers' pricing policies—how they will react to a change in price, and what that might mean for its own policy. In fact, oligopolistic pricing decisions resemble moves in a chess game. The thinking may be so complicated that no one can predict what will happen. Thus, theories of oligopolistic price determination tend to be confined almost exclusively to the short run. (In the long run, virtually anything can happen.)

The Oligopolist as Monopolist

Given the complexity of the pricing problem, the oligopolistic firm—particularly if it is the dominant firm in the market—may simply decide to behave like a monopolist (because it does have some monopoly power). Like a monopolist, Burger Bippy may simply equate marginal cost with marginal revenue (see Figure 13.3) and produce Q_o units for price P_c. Here the oligopolist's price is significantly above the competitive price level, P_c, but not as high as the price charged by a pure monopolist. (If the oligopolist

were a pure monopoly, it would not have to fear a loss of business to other producers because of a change in price.) Inefficiency in this market is slightly greater than in a monopolistically competitive market—see the shaded triangular area of Figure 13.3.

The Oligopolist as Price Leader

Alternatively, oligopolists may look to others for leadership in determining prices. One producer may assume price leadership because it has the lowest costs of production; the others will have to follow its lead or be underpriced and run out of the market. The producer that dominates industry sales may assume leadership. Figure 13.4 depicts a situation in which all the firms are relatively small and of equal size, except for one large producer. The small firms' collective marginal cost curve (minus the large producer's) is shown in part (a), along with the market demand curve, D_m. The dominant producer's, marginal cost curve, MC_d, is shown in part (b) of Figure 13.4.

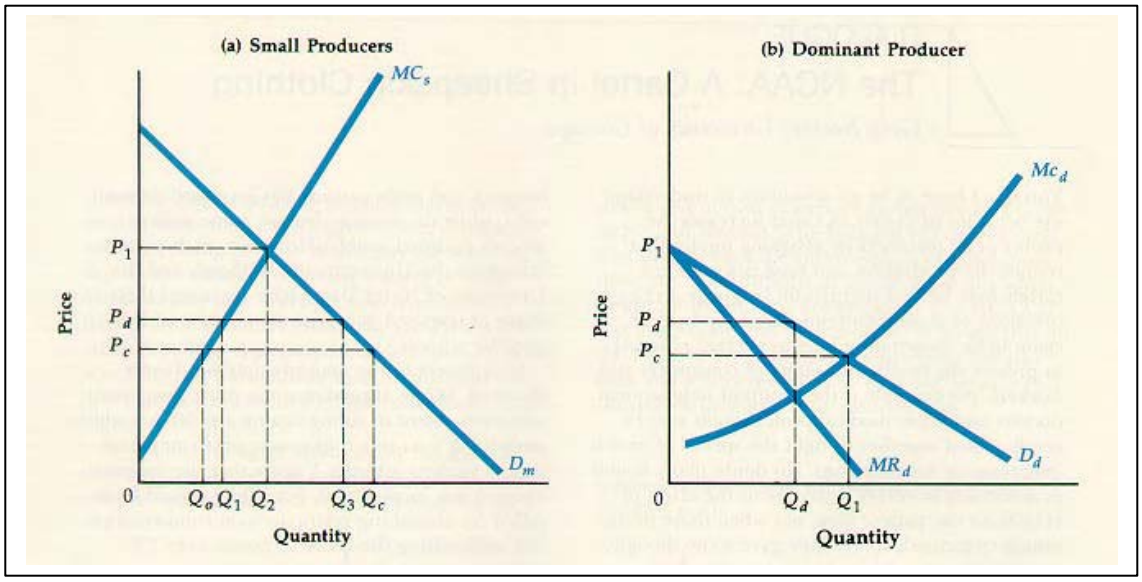


FIGURE 13.4 The Oligopolist as Price Leader

The dominant producer who acts as a price leader will attempt to undercut the market price established by small producers (part (a)). At price P_1 the small producers will supply the demand of the entire market, Q_2. At a lower price—P_d or P_c—the market will demand more than the small producers can supply. In part (b), the dominant firm determines its demand curve by plotting the quantity it can sell at each price in part (a). Then it determines its profit-maximizing output level, Q_d, by equating marginal cost with marginal revenue. It charges the highest price the market will bear for that quantity, P_d, forcing the market price down to P_d in part (a). The dominant producer sells Q_3-Q_1 units, and the smaller producers supply the rest.

The dominant producer can see from part (a) that at a price of P_1, the smaller producers will supply the entire market for the product, say, steel. At P_1 the quantity demanded, Q_2, is exactly what the smaller producers are willing to offer. At P_1 or above, therefore, the dominant producer will sell nothing. At prices below P_1, however, the total quantity demanded exceeds the total quantity supplied by the smaller producers. For

example, at a price of P_d the total quantity demanded in part (a) is Q_3, whereas the total quantity supplied is Q_1. Therefore the dominant producer will conclude that at price Pd, it can sell the difference, Q_3-Q_1. For that matter, at every price below P_1, it can sell the difference between the quantity supplied by the smaller producers and the quantity demanded by the market.

As the price falls below P_1, the gap between supply and demand expands, so that the dominant producer can sell larger and larger quantities. If these are plotted on another graph, they will form the dominant producer's demand curve, D_d (part (b)). Once it has devised its demand curve, the dominant producer can develop its accompanying marginal revenue curve, MR_d, also shown in Figure 13.4(b). Using its marginal cost curve, MC_d, and its marginal revenue curve, it establishes its profit-maximizing output level and price, Q_d and P_d.

The dominant producer knows that it can charge price P_d for quantity Q_d, because that price-quantity combination (and all others on curve D_d) represents a shortage not supplied by small producers at a particular price in part (a). Q_d, as noted earlier, is the difference between the quantity demanded and the quantity supplied at price P_d. So the dominant producer picks its price, P_d. And the smaller producers must follow.[3] If they try to charge a higher price, they will not sell all they want to sell.

Price Stability and the "Kinked" Demand Curve

Several decades ago, economists believed they had noticed something quite significant about oligopolies. For relatively long periods of time, prices in these industries seemed to remain more or less fixed. This observed "stickiness" of oligopolistic prices gave rise to the theory of the "kinked" demand curve—a theory that tries to explain not how prices are determined, but why they do not move very much.

Figure 13.5 shows the hypothetical kink in the oligopolist's demand curve that was thought to produce price stickiness. The notion was that the interdependent nature of oligopolistic pricing decisions gave rise to the kink. Suppose the price of steel is P_1. An oligopolistic firm can reason that if it lowers its price, other firms will follow suit to protect their shares of the market. Therefore, the demand curve below that point is relatively inelastic. If the firm raises its prices, however, it will lose customers to the other firms, who have no reason to follow a price increase. The demand curve above P_1 is therefore relatively elastic.

Because of the kink at P_1 the marginal revenue curve is discontinuous. At an output of Q_1, a gap develops between the upper and lower portions of the curve (see Figure 13.5). The existence of this gap is easier to understand if one thinks of the kinked demand curve as two separate curves intersecting at the kink. The curve's bottom half

---

[3] Consider market equilibrium with and without the dominant producer. In the absence of the dominant producer, the market price will be P_1, the equilibrium price for a market composed of only the smaller producers. The dominant producer adds quantity Q_d , which causes the price to fall, forcing the smaller producers to cut back production to Q_1 in part (a).

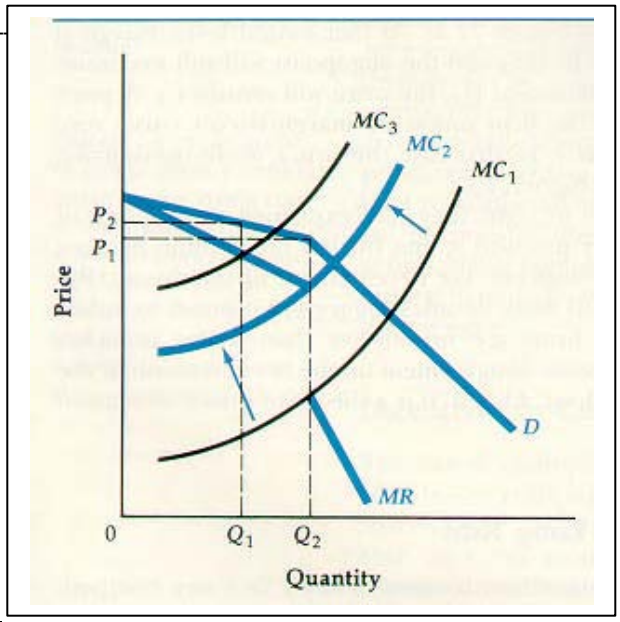


FIGURE 13.5 The Kinked Demand Curve

The theory of the kinked demand curve is based on the questionable premise that an oligopolist's prices are relatively rigid, or unresponsive to cost increases. According to the theory, the individual oligopolist reasons that other oligopolists will match a price reduction in order to protect their market shares, but will not match a price increase. The individual oligopolist's demand curve is therefore kinked at the established price: the bottom part is less elastic than the top, where even a small increase in price will cause customers to go elsewhere. Given the kinked demand curve, the firm's marginal revenue curve will be discontinuous. Even if the oligopolist's marginal cost curve shifts upward from MC_1 to MC_2, the firm will not change its price-quantity combination, P_1Q_2.

belongs to demand curve D_1 in Figure 13.6, and its top half to demand curve D_2. Seen that way, the two-part marginal revenue curve in Figure 13.5 is simply the composite of the relevant portions of the marginal revenue curves MR_1 and MR_2 in Figure 13.5. At that output level, marginal cost can shift all the way up to MC_2 and the oligopolist will still maximize profits. As long as output remains at Q_2, the price will remain P_1. A price increase would not benefit the firm unless its marginal cost curve rose higher than MC_2– say to MC_3. In that case the firm's profit-maximizing price would be only slightly higher, P_2.

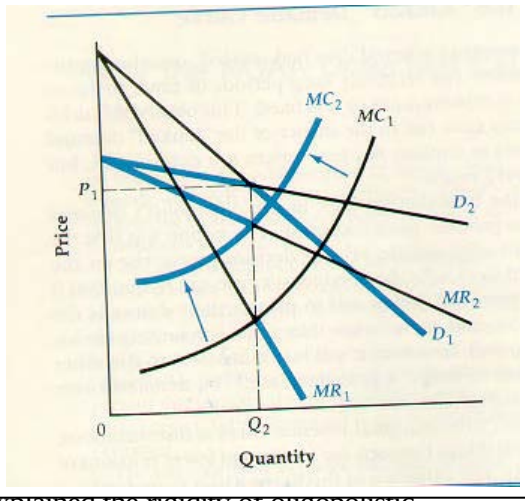


FIGURE 13.6 The Kinked Demand Curve as Two Separate Curves

The oligopolist's kinked demand curve can be viewed as the composite of two different demand curves. The portion above the kink comes from the top of a demand curve (D_2) that is relatively elastic. The portion below the kink comes from the bottom of a demand curve (D_1) that is less elastic.

Economists at one time thought they had explained the rigidity of oligopolistic prices. The only problem is that further observation has cast doubt on the evidence that motivates the development of the theory. Research conducted over the last three decades

suggests that prices in industries dominated by a few firms are no stickier than prices in other industries. Because there is some disagreement on the interpretation of the data, the theory remains with us. At best, it is a theory in search of reasonable confirmation.

The Oligopolist in the Long Run

In an oligopolistic market, significant barriers to entry face new competitors. Firms in oligopolistic industries can therefore retain their short-run positions much longer than can monopolistically competitive firms.

Oligopoly is normally associated with the automobile, cigarette, and steel markets, which include some extremely large corporations. There the financial resources required to establish production on a competitive scale may be a formidable barrier to entry. One cannot conclude that all new competition is blocked in an oligopoly, however. Many of the best examples of oligopolies are found in local markets—for instance, drugstore, stereo shops, and lumber stores—in which one, two, or at most a few competitors exist, even though the financial barriers to entry could easily be overcome. Even in the national market, where the financial requirements for entry may be substantial, some large firms have the financial capacity to overcome barriers to entry. If firms in the electric light bulb market exploit their short-run profit opportunities by restricting production and raising prices, outside firms like General Motors Corporation can move into the light bulb market and make a profit. In recent years, General Motors has in fact moved into the market for electronics and robotics.

Oligopoly power remains a cause for concern. The basis for competition, however, is the relative ability of firms to enter a market where profits can be made—not the absolute size of the firms in the industry. The small regional markets of a century ago, isolated by lack of transportation and communication, were perhaps less competitive than today's markets, even if today's firms are larger in an absolute sense. In the nineteenth century the cost of moving into a faraway market effectively protected many local businesses from the threat of new competition.

Cartels: Monopoly through Collusion

In either a monopolistically competitive market or an oligopolistic market (or even sometimes in a competitive market), firms may attempt to improve their profits by restricting output and raising their market price. In other words, they may agree to behave as it they were a unified monopoly, an arrangement called a cartel. A **cartel** is an organization of independent producers intent on thwarting competition among themselves through the joint regulation of market shares, production levels, and prices. The principal purpose of their anticompetitive efforts is to raise their prices and profits above competitive levels. In fact, however, a cartel is not a single unified monopoly, and cartel members would find it very costly to behave as if they were.

Incentives to Collude and to Cheat

The size of monopoly profits provides a real incentive for competitors to collude—to conspire secretly to fix prices, production levels, and market shares. Once they have reduced market supply and raised the price, however, each has an incentive to chisel on the agreement. The individual competitor will be tempted to cut prices in order to expand sales and profits. After all, if competitors are willing to collude for the purpose of improving their own welfare, they will probably also be willing to chisel on cartel rules to enhance their welfare further. The incentive to chisel can eventually cause the demise of the cartel. If a cartel works for long, it is usually because some form of external cost, such as the threat of violence, is imposed on chiselers.[4]

Although a small cartel is usually a more workable proposition than a large one, even small groups may not be able to maintain an effective cartel. Consider an oligopoly of only two producers, called a duopoly. A **duopoly** is an oligopolistic market shared by only two firms. To keep the analysis simple, we will assume that each duopolist has the same cost structure and demand curve. We will also assume a constant marginal cost, which means that marginal cost and average costs are equal and can be represented by one horizontal curve. Figure 13.7 shows the duopolists' combined marginal cost curve, *MC*, along with the market demand curve for the good, *D*. The two producers can maximize monopoly profits if they restrict the total quantity they produce to Q_m and sell it for price P_m. Dividing the total quantity sold between them, each will sell Q_1 at the monopoly price ($2 \times Q_1 = Q_m$). Each will receive an economic profit equal to the shaded area bounded by $ATC_1 P_m ab$, which is equal to total revenues ($P_m \times Q_1$) minus total cost ($ATC_1 \times Q_1$).

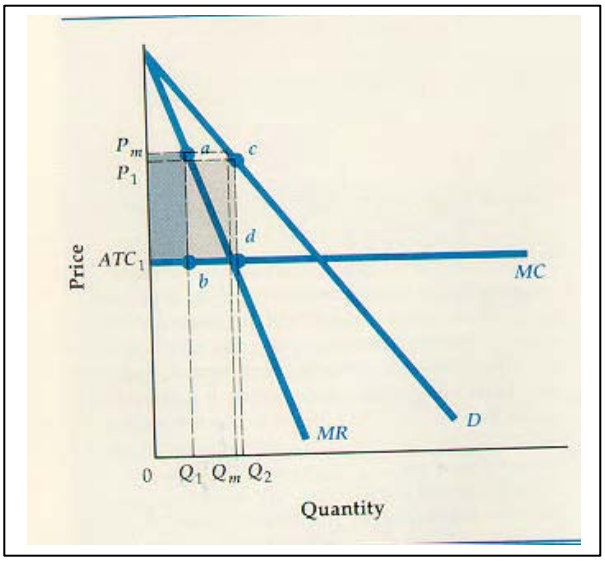


FIGURE 13.7 A Duopoly (Two-Member Cartel)

In an industry composed of two firms of equal size, firms may collude to restrict total output to Q_m and sell at a price of P_m. Having established that price-quantity combination, however, each has an incentive to chisel on the collusive agreement by lowering the price slightly For example, if one firm charges P_1, it can take the entire market, increasing its sales from Q_1 to Q_2. If the other firm follows suit to protect its market share, each will get a lower price, and the cartel may collapse.

Once in that position, each firm may reason that by reducing the price slightly -- say to P_1 -- and perhaps disguising the price cut through customer rebates or more attractive credit terms, it can capture the entire market and even raise production to Q_2.

[4] A cartel may provide members with some private benefit that can be denied nonmembers. For example, local medical associations can deny nonmembers the right to practice in local hospitals. In that case, the cost of chiseling is exclusion from membership in the group.

Each firm may imagine that its own profits can grow from the area bounded by ATC_1P_mab to the much larger area bounded by ATC_1P_mcd. This tempting scenario presumes, of course, that the other firm does not follow suit and lower its price. Each firm must also worry that the other will chisel, cut the price, and steal its market.

Thus each duopolist has two incentives to chisel on the cartel. The first is offensive, to garner a larger share of the market and more profits. The second is defensive, to avoid a loss of its market share and profits. Generally, firms that seek higher profits by forming a cartel will also have difficulty holding the cartel together, for much the same reason. As each firm responds to the incentive to chisel, the two undercut each other and the price falls back toward (but not necessarily to) the competitive equilibrium price, at the intersection of the marginal cost and demand curves. Just how far price will decline depends on the firms' ability to impose penalties on each other for chiseling.

The strength and viability of a cartel depend on the number of firms in an industry and the freedom with which other firms can enter. The larger the number of actual or potential competitors, the greater the cost of operating the cartel, of detecting chiselers, and of enforcing the rules. If firms differ in their production capabilities, the task of establishing each firm's share of the market is more difficult. If a cartel member believes it is receiving a smaller market share than it could achieve on its own, it has a greater incentive to chisel. Because of the built-in incentives first to collude and then to chisel, the history of cartels tends to be cyclical. Periods in which output and prices are successfully controlled are followed by periods of chiseling, which lead eventually to the destruction of the cartel.

Government Regulation of Cartels

Government can either encourage or discourage a cartel. Through regulatory agencies that fix prices, determine market shares, and impose penalties for violation of rules, government can keep competitors or cartel members from doing what comes naturally—chiseling. In doing so, government may be providing an important service to industry. Perhaps that is why, in most states, insurance companies oppose deregulation of their rate structures. In seeking or welcoming regulation, an industry may calculate that it is easier to control one regulatory agency than a whole group of firms plus potential competitors.

Thus in 1975, the airline industry opposed President Ford's proposal that Congress curtail the power of the Civil Aeronautics Board to set rates and determine airline routes. As the *Wall Street Journal* reported,

> The administration bill quickly drew a sharp blast form the Air Transport Association, which was speaking for the airline industry. The proposed legislation "would tear apart a national transportation system recognized as the finest in the world," the trade group said, urging Congress to reject it because it would cause "a major reduction or elimination of scheduled air service to many communities and would lead inevitably to increased costs to consumers."[5]

---

[5] "Less Regulation of Airline Sector Is Urged by Ford," Wall Street Journal, October 9, 1975, p. 3.

The real reason the airlines opposed deregulation became clear in the early 1980s, when several airlines filed for bankruptcy. Partial deregulation, begun in 1979, had increased competition, depressing fares and profits. Fares began to rise again in 1980, mainly because of rapidly escalating fuel costs. Real fares have nonetheless fallen since deregulation.

Government can suppress competition in many other ways that have nothing to do with price. Prohibiting the sale of hard liquor on Sunday, for example, can benefit liquor dealers, who might otherwise be forced to stay open on Sundays. In Florida, a state representative who managed to get a law through the legislature permitting Sunday liquor sales was denounced by liquor dealers. As one dealer commented, the legislator had "pulled the boner of the year."[6]

Cartels with Lagged Demands

Our analysis of cartels has been based on the presumption of a "standard good," one not subject to the forces of lagged demand introduced in an earlier chapter. Under market conditions of lagged demand, the pricing strategies of a cartel are potentially different. When the market is split among two or more producers, then each firm can understand that if it lowers its price, then more goods will be sold currently, but even more goods will be sold in the future, when the benefits of the lagged demand/rational addition kick in. However, each can reason that the additional future sales generated by its current price reduction could be picked up by one of the other producers. The benefits are, in other words, external. So each producer can reason that it should not incur the current costs of a lower price for the benefit of others. Each producer individually has an impaired incentive to lower the price.

On the other hand, each producer can also see that they all have a collective incentive to lower the price currently. Why? To stimulate future demand and to raise their future price and profits. A cartel under such circumstances would be organized to do what they all have an interest in doing, lower the price (not raise the price as is true in conventional markets). The problem is that the incentive to not go along with or chisel on the cartel remains strong for each firm, as is true in the conventional case, which suggest that consumers may not get the lower current price because of cartel cheating.

However, not all is lost. If firms are inclined to chisel on such a cartel, there is a potential solution that might be seen as perfectly legal by the antitrust authorities. One firm can buy the other firms simply because their profits and stock prices will be suppressed by the inability of the firm to develop a workable cartel. Once one firm controls the market, then that one firm can lower the current price for the purpose of stimulating future demand. This one firm might end up as the sole producer but might escape prosecution as a monopolist in violation of the antitrust laws (in spite of the fact that it does what a cartel of firms can't do) simply because the net effect of the buyouts is a lower price and expanded market.

---

[6] St. Petersburg Times, June 7, 1975, p. 1-B.

Antitrust Legislation

As we have seen, monopoly power often leads to market inefficiencies, or a misallocation of resources. Reductions in monopoly power should therefore improve consumer welfare. The U.S. government's antitrust policy is designed, ostensibly, to improve market efficiency by reducing barriers to entry, breaking up monopolies, and reducing the monetary benefits of conspiring to reconstruct production or raise prices. It is based on three major laws, which have been amended and modified by court decisions: the Sherman, Clayton, and Federal Trade Commission Acts.

PERSPECTIVE: **A Real-World Case of Price Fixing**

During the 1950s, General Electric Company, Westinghouse Electric Corporation, Allis-Chalmers, Southern States Equipment, and other firms and their executives were accused of conspiring to set prices and divide the market for electrical equipment. [1] Their conspiracy, which covered everything from two-dollar insulators to turbine generators, illustrates the incentives for competitors first to collude and then to chisel on their collusive agreement. As a result of a court case brought against them, which ended in 1961, fines of nearly $2 million were levied against the conspirators and the companies they represented. Six corporate executives were sent to prison, and twenty-four others were fined or given suspended sentences. It was the largest case brought to trial in the history of antitrust law, a classic example of the benefits and pitfalls of industrial conspiracy.

The seeds of the conspiracy were planted during the Second World War, when the prices of various types of electrical equipment were regulated by the Office of Price Administration (OPA). Under the auspices of the National Electrical Manufacturers Association, firms met on a regular basis to determine how they could supply the heavy wartime demand for electrical equipment. After their meetings, executives would regroup to talk about how they could get the OPA to raise prices.

When the war was over and prices were no longer controlled, these manufacturers faced competition from a growing number of smaller companies. Increasingly, buyers were asking for sealed bids as a means of getting the lowest possible prices. The major manufacturers continued their meetings, this time to talk about price fixing and methods of dividing the market. They decided to agree on their bids ahead of time and to rotate the privilege of making the lowest bid. After learning what the lowest bid would be, the others would make higher bids. The business was divided on the basis of past sales volume. In the circuit-breaker market, for example, General Electric received 45 percent of the business, Westinghouse 35 percent, Allis-Chalmers 10 percent, and Federal Pacific 10 percent.

For a more detailed account of this case, see Richard Austin Smith, "The Incredible electrical Conspiracy," parts I and II, *Fortune*, April and May 1961, pp. 132 ff. (April) and pp. 161 ff. (May).

The Sherman Act

The Sherman Act was passed in 1890, after a series of major corporate mergers. It contains two critical provisions. The first, Section 1, declares illegal "every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among several states or with foreign nations." The second, Section 2, declares that "every person who shall monopolize, or conspire with any other person or persons to monopolize any part of trade or commerce among the several states, or with foreign nations, shall be guilty of a misdemeanor. . . ." In short, the first section outlaws any form of cooperative behavior that restrains competition; the second outlaws monopolization or any attempt to acquire monopoly power.

The language seems clear enough, yet the courts were initially reluctant to rule against violations of the law, citing prosecutors' loose interpretation of the words "restraint of trade" and "conspire. . . .to monopolize." In 1911, however, the Supreme Court ruled that Standard Oil Company, which then controlled 90 percent of the nation's refinery capacity, should be broken up. By dividing the firm along geographical lines (which explains the names Standard Oil of Ohio and Standard Oil of California), the court effectively nullified the economic benefits of the breakup. In place of one large monopoly, the justices created smaller monopolies. Later, the court broke up the United States Steel Corporation and American Can Company on the grounds that they and followed "unfair and unethical" business practices.

The Clayton Act

Because the Sherman Act did not specify what constituted unfair and unethical business practice, and because the courts generally took a very narrow view of what constituted restraint of trade and commerce, Congress passed a new law in 1914. The Clayton Act listed four illegal practices in restraint of competition. It outlawed price discrimination, or the use of price differences not justified by cost differentials to lessen competition or create a monopoly. This provision was intended to prevent firms from cutting prices below cost in a particular geographical region in order to drive competitors out of the market. Railroads and department stores were allegedly involved in such "predatory competition."

The Clayton Act also forbade tying contracts and exclusive dealerships. A **tying contract** is an agreement between seller and buyer that requires the buyer of one good or service to purchase some other product or service. If IBM tried to force buyers of home computers to purchase only IBM software, for example, its purchase and sale agreement with customers might be considered a typing contract. An **exclusive dealership** is an agreement between a manufacturer and its dealers that forbids the dealers from handling other manufacturers' products. The Clayton Act is applicable only to exclusive dealerships that reduce competition "substantially," however. As long as other manufacturers' products are sold in the same area, manufacturers may organize exclusive dealerships covering designated territories, as is common in the automobile industry. Since 1985, the antitrust enforcement agencies and the courts have been more lenient toward such nonprice vertical restraints as tying contracts and exclusive dealerships.

Section 7 of the Clayton Act forbids mergers, or the acquisition by a firm of its competitors' stock, if the effect of the merger is to reduce competition substantially. The act applies only to horizontal mergers, however.

A **horizontal merger** is the joining of two or more firms in the same market—for example, two car companies--into a single firm. Vertical mergers were excluded from the act. A **vertical merger** is the joining of two or more firms that perform different stages of the production process into a single firm. For example, the Clayton Act would permit the merging of an oil-drilling firm with a refining firm. So would conglomerate mergers. A **conglomerate** is a firm that results from the merging of several firms from different industries or markets. The combining of firms in tow entirely different markets—washing machines and light bulbs, for instance,--would be considered a conglomerate merger. These loopholes in the Clayton Act—vertical and conglomerate mergers—were closed in part by the Celler-Kefauver Antimerger Amendment, passed in 1950. Although the act has since been applied to vertical mergers, it has never been applied to conglomerates.

Finally, the Clayton Act declared interlocking directorates illegal. An **interlocking directorate** is the practice of having the same people serve as directors of two or more competing firms. If the same people direct competing firms and advise policies that effectively reduce industry output, they constitute a defacto monopoly. Section 8 of the Clayton Act prohibits such arrangements if they "substantially reduce" competition.

The Federal Trade Commission Act

The original purpose of the Federal Trade Commission Act, passed in 1914, was to thwart "unfair methods of competition" among firms. The act empowered the Federal Trade Commission to investigate cases of industrial espionage, bribery for the purpose of obtaining trade secrets or gaining business, and boycotts.[7] Later the Wheeler-Lea Amendment expanded the commission's mandate to cover "unfair or deceptive acts or practices" that harmed customers, including the sale of shoddy merchandise and misleading or deceptive advertising.

The Purposes and Consequences of Antitrust Laws

The ostensible purpose of all these laws is to fight monopoly power by outlawing business practices that prevent or retard competition. By forcing firms to restrict production or fix prices surreptitiously, antitrust legislation makes collusion among competitors more costly. Violations of the law carry fines and penalties on conspiring firms and their employees.

---

[7] Not all boycotts are prohibited, of course—only efforts designed to prevent goods from reaching their intended designation. That is, a union cannot prevent goods from crossing its picket lines, and firms cannot organize restrictions on the purchase of other firms' products.

PERSPECTIVE: **Economic Consequences of Treble Damages**

Section 4 of the Clayton Act states that

> Any person who shall be injured in his business or property by reason of anything forbidden in the antitrust laws may sue therefore in any district court of the United States in the district in which the defendant resides or is found or has an agent, without respect to the amount in controversy, and shall recover threefold the damages by him sustained, and the cost of suit, including a reasonable attorney's fee.

In other words, the successful private plaintiff in an antitrust case is to be paid treble the damages done to him by the defendant. This provision of the Clayton Act means that thousands of private firms and individuals join the Department of Justice and the Federal Trade Commission in the enforcement of antitrust laws. For many years the treble damages provision generated no controversy; it accorded well with the notion that victims should be compensated and apparently served an important deterrent to potential violators. Beginning in the 1970s, criticism of treble damages began to appear in the law and economics literature.

Critics have pointed out that the law has costs as well as benefits. A proper assessment must take account of both costs and benefits. Economists William Breit and Kenneth G. Elzinga find three principal costs of treble damage suits: the perverse incentives effect, the misinformation effect, and reparations costs.

Perverse incentives. Treble damages can reduce the incentives of consumers to take private steps to avoid the harm done by the monopolistic firm. If the expected gains from the successful antitrust suit are high relative to the costs of buying from a monopoly seller, the buyer has a positive incentive not to avoid the monopoly seller, even if it is possible to do so. To put it another way, the treble damage provision encourages private enforcement of antitrust laws but discourages the private prevention of monopoly behavior.

Misinformation. A private party has an incentive to claim damages from anticompetitive behavior even when such behavior has not taken place. The treble damages provision generates many "nuisance suits" in which the plaintiff sues in the hope of forcing an out-of-court settlement. Such tactics have a fair chance of success in antitrust cases because in many instances the definition of anticompetitive behavior is quite vague. Moreover, in a jury trial anything can happen, giving the defendant (even if innocent) a strong incentive to settle before going to court.

Reparations costs. Considerable resources are devoted to determining and allocating damages in private antitrust suits. The judicial, clerical, and legal costs associated with compensating private plaintiffs all represent costs incurred solely because of the private enforcement provisions of the antitrust laws.

Although treble damages have its defenders, many students of law and economics have suggested that the provision be done away with. Richard Posner and others have suggested reducing private antitrust claims to single damages. Others have supported severely limiting the types of cases subject to the treble damage provision. Elzinga and Breit support pure public enforcement of the antitrust statutes.

The courts themselves seem to have grown wary of treble damages. Judges have in several recent rulings reduced damage awards in treble damage cases. The behavior of the judges in such cases may reflect the belief that the broad application of the treble damage provision generates more costs than benefits to the economy.

Although many economists believe antitrust law has achieved some of these objectives, critics complain of its inefficiency. Detecting violators and bringing legal action against them takes time. Often market forces erode monopoly power before the government can prosecute. The result can be a huge waste of legal resources. As noted in the last chapter, the Department of Justice spent over twelve years prosecuting IBM for its dominance in the mainframe computer market, with questionable results.

In attempting to determine which firms possess monopoly power, the Department of Justice and the Federal Trade Commission have sometimes relied on "concentration ratios," or estimates of the percentage of industry sales controlled by the largest domestic firms. The arbitrary use of such ratios can be misleading. The top four firms in steel, for example, may have little monopoly power, for they must compete with producers of fiberglass, aluminum, and wood as well as with each other. Moreover, large market shares may be the result of superior efficiency, a higher quality product, or good luck. Nevertheless, to avoid the appearance of impropriety, firms may decide to operate on a smaller scale than competitive principles would normally dictate.

Finally, as amended, the Clayton Act gives private firms the power to initiate antitrust suits. If an antitrust violation is proven, prosecuting firms receive a reward equal to three times the computed damages. Critics charge that firms may use antitrust suits as a means of diverting their competitors' resources away from production. The mere threat of an antitrust suit may be enough to keep some large firms from competing actively through better product design and lower prices.

MANAGER'S CORNER: **"Hostile" Takeover as a Check on Managerial Monopolies**

It may appear that our discussion of monopolies applies only to "markets," and has little or nothing to do with the management of firms. Indeed, the theory of monopolies is directly applicable to management problems. This is because firms often rely exclusively on internal departments (and their employees) for the provision of a variety of services, legal, advertising, accounting, as well as the production of parts that are assembled into the firm's final goods sold to consumers. In such cases, the internal departments can begin to act like little monopolies, cutting back on what they could produce and demanding a higher price (through their firm's budgetary processes) for what they do than is required. One reason a firm might want to outsource its services is that it does not become controlled by internal monopolies; the firm can always seek competitive bids from alternative outside suppliers. The act of outsourcing some services can also keep the outsourcing threat alive for other internal department that might try to act like an internal monopoly.

Still, managers can become complacent in managing their departments, allowing their departments to act monopolistically – and inefficiently. Corporate takeovers, however, represent an important check on management discretion, and on the extent to which internal departments can behave monopolistically.

Reasons for Takeovers

There are many reasons for corporate takeovers and different ways for them to occur. There may be complementarities in the production and distribution of the products of two firms that can be best realized by one firm. For example, Disney produces programs that can be aired on ABC's TV network as well as company owned stations. Or, as was commonly the case in earlier manufacturing mergers, two firms may find that they can realize economies of scale by combining their operations. And one firm may be supplying another firm with the use of highly specific capital and a merger between the two reduces the threat of opportunistic behavior that can be costly to both.

Most takeovers are what are referred to as "friendly." A friendly takeover occurs when the management of the two firms works out an arrangement that is mutually agreeable. The takeover of ABC by Disney was a friendly one. Indeed, takeovers occur for the same reason all market transactions occur: Generally speaking, efficiencies are expected, meaning that both parties can be made better off. So it should not be surprising that most takeovers are friendly.

But there are takeovers that are opposed by the management of the firms being taken over, as was the case, at least initially, in IBM's takeover of Lotus. These takeovers are referred to as "hostile" and are commonly seen as undesirable and inefficient. "Hostile" takeovers are depicted as the work of corporate "raiders" who are only interested in turning a quick profit, who disrupt productivity by forcing the management of the targeted firms to take expensive defensive action and distracting them from long-run concerns.

If managers of target corporations always act in the interest of their shareholders (the real owners of the corporation), then a strong case could be made that so-called hostile takeovers are inefficient. Managers of the target corporation would then oppose a takeover only if it could not be made in a way that benefited their shareholders, as well as those of the acquiring corporation. But if managers could always be depended upon to act in the interest of their shareholders, then there would be no need for many of the corporate arrangements that have been discussed in this book.

Indeed, the strongest argument in favor of "hostile" takeovers is that they bring the interests of managers more in line with those of shareholders than would otherwise be the case. There is a so-called "market for corporate control" that allows people who believe that they can do a better job managing a company and maximizing shareholder return, to oust the existing management by outbidding them for the corporate stock. Although there are not a large number of such takeover attempts, and not all attempts are successful, just the threat of a "hostile" takeover provides a strong disincentive for managers to go as far as they otherwise would like in pursuing personal advantages at the expense of their shareholders. This suggests that there are efficiency advantages from "hostile" takeovers, a proposition that is much debated. The issue of efficiency is not unrelated, however, to the primary concern of this chapter, which is why "hostile" takeovers are less hostile than they are commonly depicted.

A takeover is often considered hostile for the very reason that it promotes efficiency. A management team that is doing a good job managing a firm efficiently has

little to fear from being taken over by a rival management team. The stock price of a well-managed firm will generally reflect that fact, and it will not be possible for a corporate raider to profit by buying that firm's stock in the hope of increasing its price through improved management. Only when the existing managers are not running the firm efficiently, either because of incompetence, the inability to abandon old ways in response to changing conditions, or by intentionally benefiting personally at the expense of shareholders, is a takeover likely. But under these circumstances, a takeover that promises to increase efficiency will not be popular with existing managers since it puts them out of work. Not surprisingly, managers whose jobs are threatened by a takeover will see it as "hostile."

The fact that pejorative terms such as "hostile takeover" and "corporate raiders" are so widely used is testimony to the advantage existing managers have over shareholders at promoting their interests through public debate. The costs from a "hostile" takeover are concentrated on a relatively small number of people, primarily the management team that loses its pay, perks, and privileges. Each member of this team will lose a great deal if the team is replaced and so has a strong motivation to oppose a takeover. And even a grossly inefficient management team can be organized well enough to respond in unison to a takeover threat, and to speak in one voice. That voice will usually characterize a takeover as hostile to the interests of the corporation, the shareholders, the community, and the nation, and we might expect managers to be more vociferous the more inefficient the management.

But if a takeover is actually efficient, what about the voice of those who benefit? Why is the media discussion of takeovers dominated by the managers who lose rather than by the shareholders who win? And there is plenty of evidence that the shareholders of the target company in a hostile takeover do win. For example, during the takeover wave in the 1980s, it has been estimated that stock prices of targeted firms increased about 50 percent because of a hostile takeover, which suggests that the managers of the targeted firms may have destroyed a considerable amount of their corporations' value before being targeted for takeover.[8] As will be discussed later, this increase in stock values does not necessarily *prove* that a takeover is efficient. The stock prices of the firm that is taking over the target firm could be depressed, for example.[9] But even if the takeover is not efficient, the shareholders of the target firm should favor it and counter the negative portrayal put forth by their managers. This seldom happens, however, because there are typically a large number of shareholders, with few, if any, having more than a relatively small number of shares. Most shareholders have a diversified portfolio

---

[8]See Michael C. Jensen, "Takeovers: Their Causes and Consequences," Journal of Economic Perspectives, vol. 2, no. 1 (Winter 1988), pp. 21-48.

[9]However, Michael Jensen minces few words on what the data implies, "[T]he fact that takeover and LBO premiums average 50% above market price illustrates how much value public-company managers can destroy before they face a serious threat of disturbance. Takeovers and buyouts both create value and unlock value destroyed by management through misguided policies. I estimate that transactions associated with the market for corporate control unlocked shareholder gains (in target companies alone) of more than $500 billion between 1977 and 1988 – more than 50% of the cash dividends paid by the entire corporate sector over this same period" [Michael C. Jensen, "Eclipse of the Public Corporation," Harvard Business Review (September-October 1989), pp. 64-65].

and are only marginally affected by changes in the price of any particular corporation's stock. The probability that the actions of a typical individual stockholder will have an impact is very low, approaching zero. So even if the gain to shareholders far exceeds the loss to management, the large number of shareholders and their diverse interests make it extraordinarily difficult for them to speak in unison. Shareholders are not likely to influence the terms of the debate in ways that promote their collective interest.

If shareholders and management were on equal footing at influencing the public perception of hostile takeovers, almost no takeovers would be reported as hostile. Consider a hypothetical situation that is similar to what is commonly seen as a hostile takeover.

Assume that you are the owner of a beautiful house on a high bluff overlooking the Pacific Ocean near Carmel, California. You are extremely busy as a global entrepreneur and unable to spend much time at this house. Since the house and grounds require full-time professional attention, you have hired a caretaker to manage the property. Assume that you pay the caretaker extremely well (mainly because you want him to bear a cost from being fired for shirking and engaging in opportunism), and give him access to many of the amenities of the property. He's very happy with the job, and you are pleased enough with his performance.

But one day a wealthy CEO who is planning to retire in the Carmel area makes you an offer on the house of $15 million, about 50 percent more than you thought you could sell it for. Although you were not interested in selling at $10 million, you find the $15 million offer very attractive. For whatever reason, the house is worth more to the retiring CEO than to you. It could be that the CEO values the property more than you simply because she will have more time to spend living in and enjoying the house. Or it could be because the CEO believes a profit can be made on the house by bringing in a caretaker who will do a far better job managing the property, thus increasing its value to above $15 million. But it really makes little difference to you why the CEO values the house more than you do, and you are quite happy to sell at the price offered whatever the reason.

Imagine how surprised you would be if, as the sale of your house was being negotiated, the news media reported that your property was the target of a hostile takeover by a "house raider" only interested in personal advantage. What's so hostile about being offered a higher price for your property than you thought it was worth? And are you somehow worse off because the buyer also sees private benefit in exchange?

But the media wasn't interested in your opinion. Instead, reporters had been talking to your caretaker who knew he would lose his job if the sale went through. So the caretaker was reporting that the sale of the property was the result of a hostile move by an unsavory character. Obviously this is silly, and the media is not likely to report this, or any similar sale of a house, as a hostile takeover. But is this any sillier than reporting a corporate takeover as hostile when the owners of the corporation (the shareholders) are being offered a 50 or 100 percent premium to sell their shares? Not much.

The two situations are not exactly the same, but they are similar enough to call into question the hostility of most hostile takeovers. One important difference between

the two situations is that if such a report did start to circulate about the sale of your house, and somehow threatened that sale, you would have the motivation and ability to clearly communicate that it was your house, that you found the offer attractive, and that there was nothing at all hostile about the sale. This difference explains why our example should not be taken as a criticism of the press. When there is one owner (or a few), as in the case of a house, the press can easily understand and report that owner's perspective. But when there are thousands of owners, as in the case of corporations, it is much easier for reporters to obtain information about a corporation from its top managers.

The fact that there are a multitude of owners in the case of corporations is the basis for other differences between the sale of a house and the sale of a corporation. Just as reporters find that it is easier to rely on top management for information on a corporation, so do the owners of a corporation find it easier to rely on management to make most corporate decisions, even major decisions such as those that affect the sale of the corporation. Obviously, the reason for granting a management team the power to act somewhat independently of shareholders is that shareholders are so large in number, so dispersed in location, and so diverse in interests, that they cannot make the type of decisions needed to manage a corporation, or much of anything else for that matter. But as we have discussed in detail throughout this book, there are risks associated with letting agents (managers) act on behalf of principals (owners/shareholders). As the owner of the house outside Carmel, would you want your caretaker to negotiate the sale for you? Only if the caretaker were subject to a set of incentives that go a long way in lining up his interests in the sale with yours.

The reason many corporate practices and procedures are what they are can often be explained in terms of motivating corporate agents to behave in ways that serve the interests of their principals. Aligning the interests of managers with those of owners when there are attempts by outsiders to gain control of a corporation from the current management team is particularly difficult. There are corporate arrangements (to be considered later), however, that are best understood as motivating corporate managers to take shareholders' interests into account in the case of takeover offers. These arrangements aren't perfect, as evidenced by the popularity of the terms "hostile takeover" and "corporate raider." It should be emphasized though that both shareholders and managers can benefit from such arrangements.

The benefit to shareholders from arrangements that motivate a management team to promote the stockholders' interests should be obvious. The benefit to managers is subtler. Managers who accept restrictions that reduce their ability (or incentive) to frustrate attempts by outsiders to take control of the corporation are worth more than managers not subject to such restrictions. How much would you be willing to pay an agent who could gain at your expense with impunity? So while managers can be expected to take advantage of allowable opportunities to protect their jobs against a takeover attempt, they would not want to work for a corporation that didn't go a long way to restrict those opportunities.

The most important way managers can protect themselves against a hostile takeover is by doing a good job managing. Being a good manager requires more than the skills that can be learned in an MBA program and honed with experience. It also requires

corporate arrangements that provide strong incentives for managers to work together as a team for the good of the shareholders, and that provide them with clear information on how well they are doing. These arrangements take many forms, and they are very attractive to managers quite apart from their ability to improve managerial performance. For example, few managers complain about executive compensation packages that increase in value when the price of the corporate stock increases. A corporate executive who receives a large payoff from exercising a stock option provided by his or her compensation package will tell you that income is justified because increasing stock prices reflect, at least in part, the requisite skill at making decisions that benefited the shareholders.

There is a lot of truth in this justification for high incomes for corporate managers. Although it is obviously possible for stock prices to increase or decrease for reasons having nothing to do with the performance of managers, good management decisions do have positive effects on the price of a corporation's stock. But managers who want to take some of the credit, and reward, when the corporate stock is going up should also be prepared to accept some of the consequences when the stock is going down. From the perspective of efficient incentives, it is best if managers suffer more loss from declining stock prices when they are to blame for that decline than when they are not. Though not perfect, this is what hostile takeovers tend to do. If a corporation's stock price declines because of a general decline in the stock market, or for reasons that have nothing to do with the performance of management, there is little for a "corporate raider" to gain from a takeover. The threat of a takeover, particularly a takeover existing management sees as hostile, is likely only when those mounting the takeover bid believe that better management can increase the value of the stock.

So far, we have explained why corporate takeovers that enrich the owners are often characterized as hostile in the press. The shareholders typically see nothing hostile about these takeovers, but the corporate managers whose jobs are threatened do. And managers, not shareholders, are the ones reporters turn to when they are looking for a corporate spokesperson. We have also noted that hostile takeovers are efficient for the very reason that managers consider them to be hostile: they force managers to either manage the corporation in the best interests of the shareholders or lose their lucrative jobs. The management team that is incompetent, or complacent, or that becomes more concerned with its privileges and perks than with running a tight ship, reduces the profitability of the corporation and the price of the corporate stock. This creates the opportunity for an individual, or group of individuals, to purchase the corporation's stock at a low price, take a controlling interest in the corporation, and then profit by putting in a management team whose superior performance increases the price of the stock.

The Efficiency of Takeovers

But are hostile takeovers efficient? Not everyone believes they are. Hostile takeovers are commonly seen as ways to increase the wealth of people who are already rich at the expense of the corporation's average workers (not just its managers), the corporation's long-run prospects, and the competitiveness of the general economy. For example, responding to a hostile takeover bid for Chrysler Corporation by Kirk Kerkorian, a major

newspaper editorialized, "[W]hen Kerkorian was complaining about insufficient return to stockholders, the value of [his] investment in Chrysler had more than tripled, to $1.1 billion. That's not good enough? To satisfy his greed, Kerkorian seems prepared to endanger the jobs of thousands of Americans and the health of a major corporation so important to the economy. . ."[10]

This editorial comment ignores the efficiency effects of a corporate takeover. But at the same time, the effect of a hostile takeover on economic efficiency is more complicated than has been suggested in this chapter so far. The stockholders of the corporation being taken over do gain. But what about the stockholders and bondholders of the corporation doing the taking over? Don't they lose as their firm runs up lots of debt to pay high prices for the stock of the acquired firm? Also, doesn't the threat of a hostile takeover motivate managers to make decisions that boost profits in the short run but which harm the corporation's long-run profitability? And what about the fact that important parts of an acquired firm are often spun off after a hostile takeover, leaving a much smaller firm, with many of its workers being laid off? Shouldn't these losses be set against any gains that the shareholders of acquired firms receive, and isn't it possible that the losses are larger than the gains?

These are good questions, and deserve serious consideration. But first, let's consider in more detail the magnitude of the gains to the shareholders of a corporation that is targeted for a takeover. The evidence suggests that they are quite large. For example, a study by the Office of the Chief Economist of the Securities and Exchange Commission looked at 225 successful takeovers from 1981 through 1984 and found that the average premium to shareholders was 53.2 percent. In a follow up study for 1985 and 1986 the premium was found to have dropped to an average of 37 and 33.6 percent respectively. These averages probably understate the gains because they compare the stock price one month before the announcement of a takeover bid with the takeover price, and often the price begins increasing in response to rumors long before a formal offer is tendered.[11] These percentages represent huge gains in total dollars, amounting to $346 billion over the period 1977-86 (in 1986 dollars), according to one study.[12]

Those who own something that others are bidding for should be expected to see their wealth increase. So it is not really surprising that takeover bids increase the wealth of the corporation's stockholders. But that is not necessarily true for the stockholders of a corporation mounting a takeover bid. In a competitive bidding process it is possible to bid too much, and some believe that this is particularly true of the one making the winning bid. The winning bid is typically made by the bidder who is most optimistic

---

[10] See "Long-term Risk" (editorial), Atlanta Journal and Constitution, April 15, 1995: p. A-10.

[11] Unless otherwise noted, the studies cited are discussed in Gregg A. Jarrell, James A. Brickley, and Jeffrey M. Netter, "The Market for Corporate Control: The Empirical Evidence Since 1980," Journal of Economic Perspectives (Winter 1988), pp. 49-68.

[12] See page 21 of Michael C. Jensen, "Takeovers: Their Causes and Consequences," Journal of Economic Perspectives (Winter 1988), pp. 21-48. It should be pointed out that this estimate applied to all mergers and acquisitions, not just "hostile" takeovers. But "hostile" or not, takeovers consistently increase the value of the acquired firm's stock, and probably increase it more when the takeover is opposed by management than otherwise, since offering a higher price is a way around a reluctant management.

about the value of the object of the bidding.[13] This is no problem when bidding for something the bidder wants for its subjective value (say an antique piece of furniture), since the object probably is worth more to the winning bidder than to others. But when bidding for a productive asset (such as an offshore oil field) valued for its ability to generate a financial return, the value of the object is less dependent on who owns it.[14] Therefore, if the average bid is the best estimate of the value of the object, then there is a good chance that the winning bid is too high.

Economists have referred to this possible tendency to overbid as the "Winner's Curse." But the Winner's Curse may not be all that prevalent for two very good reasons. First, people who are prone to fall victim to this curse are not likely to acquire (or retain) the control over the wealth necessary to keep bidding on valuable property, certainty not property as valuable as a corporation. Second, in many bidding situations each bidder often receives information on how much others are willing to pay as the bidding process takes place, and adjusts his evaluation of the property accordingly. This is the case in corporate takeovers where offers to pay a certain price for a corporation's stock are made publicly.

So, we should expect that the winning bid for the stock of a corporation targeted for a takeover will fairly accurately reflect the value of that corporation to the winner, and therefore not greatly affect the wealth of the acquiring corporation's stockholders, and the more competitive the bidding process, the closer the bid price to the actual stock value. And that is exactly what the evidence suggests. According to a 1987 study by economists Gregg Jarrell and Annette Poulsen, stockholders of acquiring corporations realized an average gain of between 1 and 2 percent on 663 successful bids from 1962-1985. Interestingly, and not surprisingly, as takeover activity increased, the return to acquiring firms decreased, with the average percentage return being 4.95 in the 1960s, 2.21 in the 1970s, and -0.04 (but statistically insignificant) in the 1980s.[15]

What about the possibility that the additional value realized by shareholders of the target corporation is paid for by losses to bondholders? For example, a takeover could increase the risk that either the acquiring or the acquired firm suffers financial failure, while increasing the possibility that one or both experience very high profits. Shareholders stand to benefit from the high profits if they occur, and so can find the expected value of their stock increasing because of the increased risk. The additional risk cannot generate a similar advantage from bondholders since the return to bondholders is fixed. They lose if the corporation goes bankrupt, but don't share in any increased profits if the corporation does extremely well. According to several studies of takeovers from

---

[13] See Richard H. Thayer, The Winner's Curse: Paradoxes and Anomalies of Economic life (New York: Free Press, 1992).

[14] In general, of course, the value of the asset will depend to some degree on who owns it. The highest bidder will likely have good reason to believe that he or she is better able to utilize the asset to create value. In the case of an oil field, the possibilities for one owner to obtain more wealth than another are probably quite limited. In the case of a corporation, the importance of management no doubt provides more opportunity for some owners to run the business more profitably than others.

[15] Jarrell, Gregg A., and Annette B. Paulsen, "The Returns to Acquiring Firms in Tender Offers: Evidence from Three Decades," Financial Management, vol. 18 (Autumn 1989), pp.12-19.

the 1960s into the 1980s, however, takeovers do not impose losses on bondholders.[16] No doubt some bondholders suffer small losses, while some realize small gains, but the best conclusion is that under even the worse case any losses to bondholders do not come anywhere close to offsetting the gains to stockholders.

So far we have been talking about the average wealth effect on shareholders and bondholders from takeovers. Just because the average wealth effect of a hostile takeover is positive does not mean that all such takeovers create wealth. People make mistakes in the market for corporate takeovers just as they do in other markets, and in all aspects of life. The question is not whether people make mistakes, but whether they are subjected to self-correcting forces when they do. The bidders subject to the winners curse should themselves be the subject of a takeover. The evidence suggests that in the case of hostile takeovers, they are. In a 1990 study, economists Mark Mitchell and Kenneth Lehn asked, "Do Bad Bidders Become Good Targets?" Looking at takeovers over the period January 1980-July 1988, they found that those firms resulting from takeovers that were wealth reducing (according to the response of stock prices) were more likely to be challenged with a subsequent takeover than were those takeovers that were wealth increasing. The market for corporate control does not prevent mistakes from being made, but it creates the information and motivation vital for correcting them when they occur.[17]

If you are a corporate manager you may be thinking that the threat of a takeover could motivate you to act in ways that increase the value of the corporate stock in the short run, but which are harmful to the profitability of the corporation in the long run. Is it true that managers are less likely to be ousted in a hostile takeover if they concentrate on short-run profits at the expense of long-run profits? The answer might be yes if the prices of corporate stock reacted only to short-run profits. But there is plenty of evidence indicating that stock prices reflect the market's collective estimate of the long-run profitability of corporations.[18]

People's view of the future is always cloudy and uncertain, and no one argues that stock prices are a completely accurate gauge of the present value of a corporation's future prospects. But as soon as new information becomes available on a corporation's future profitability, it is in the interest of investors to interpret this information as accurately as possible, and make decisions on the purchase or sale of stock that quickly cause the price of that stock to reflect the new information. Errors are always being made, but the errors of some create profitable opportunities for others to correct those errors with their buying and selling decisions. And those who consistently make errors soon find themselves lacking the resources (and also the desire) to continue making decisions that affect stock prices.

---

[16] Debra K. Dennis, and John J. McConnell, "Corporate Mergers and Security Returns," Journal of Financial Economics, 1986, 16, 143-187; and Kenneth Lehn and Annette B. Paulsen, "Sources of Value in Leveraged Buyouts" in Public Policy Towards Corporate Takeovers (New Brunswick, N.J.: Transaction Publisher, 1987).

[17] See Mark L. Mitchell and Kenneth Lehn, "Do Bad Bidders Become Good Targets?" Journal of Political Economy vol. 98, no. 2 (April 1990), pp. 372-398.

[18] More accurately, stock prices tend to reflect the discounted present value of the stream of profits the corporation is expected to generate.

Consider a decision facing you as a manager on whether to commit to an expensive research and development project that will reduce profits over the near term but which is expected to more than offset this loss with higher profits in the future. Should you be fearful that investing in this project will, because of the reduction in current profits, drive the price of your stock down, making your corporation more vulnerable to a hostile takeover? The answer is probably no, especially if your estimate of the long-run profitability of the R&D project is correct. There are two good reasons for believing this. First, the obvious fact that price-earnings ratios vary widely between different stocks provides compelling evidence that stock prices reflect more than current profits. Second, studies indicate that a corporation's stock price generally increases when the corporation announces increased spending on investment, and generally decreases when a reduction in investment spending is announced.[19] A study by Brownyn Hall found that, over the period 1976-85, the firms taken over by other firms did not have a higher R&D to sales ratio than did firms in the same industry that were not taken over.[20] There is no reason for managers to become short sighted because of the threat of a hostile takeover. Indeed, the best protection against a takeover, hostile or otherwise, is to make decisions that increase the long-run profitability of the corporation, even if those decisions temporarily reduce profits.

What about the fact that once a corporation is taken over it is sometimes broken up as the acquiring firm sells off divisions, often profitable divisions? Isn't this disruptive and inefficient? There is no doubt that takeovers are disruptive, particularly when they result in parts of the acquired firm being spun off. But disruption is not necessarily inefficient. Indeed, any economy that hopes to be efficient has to motivate rapid responses to changing circumstances, and those responses are necessarily disruptive. Making the best use of resources in a world of advancing technologies, improved opportunities, and global competition requires continuous disruption. The alternative is stagnation and relative decline.

Many of the mergers that took place in the 1960s and 70s created large conglomerate structures that, even if efficient at the time, soon ceased to be efficient. Increased global competition began rewarding smaller firms with quicker response times to changing market conditions. Technology reduced the synergies that might have existed at one point by having different products produced within the same firms. It became less costly for firms to buy inputs and components from other firms, thus increasing the ability to specialize in their core competencies (in the vernacular of earlier chapters, transaction costs fell).

In many cases these changes made the divisions of the corporation worth more as separate firms than as parts of the whole. Many managers, however, prefer to be in charge of a large firm than a small one and are reluctant to divest divisions that are worth more by themselves or as part of another organizational structure. This extant managerial reluctance of the 1960s, 1970s, and into the 1980s was partly responsible for the depressed stock prices that corporate raiders were able to take advantage of by buying a

---

[19] John J. McConnell and Chris J. Muscarella, "Capital Expenditure Decisions and Market Value of the Firm," Journal of Financial Economics, vol.14 (1985), pp. 399-422.

[20] Hall's study is discussed by the Jensen article cited in footnote 3.

controlling interest in conglomerates and then increasing their total value by spinning off some of their divisions.[21]

Another complaint about the spinning off of divisions and downsizing that often accompanies takeovers is that workers are laid off. The claim is made that while stockholders may come out ahead, they do so at the expense of workers who lose their jobs. There is evidence that hostile takeovers do result in reductions in the work force. But the questions we want to consider are the following

- Is this a valid criticism of takeovers?
- Which workers are most likely to be laid off and how big is the cost to the workers when compared against the gain to shareholders?

The fact that workers are laid off after hostile takeovers is consistent with the view that these takeovers promote efficiency. The most natural thing in the world for managers to do when sheltered against the full rigors of competition is to let the workforce grow larger than efficiency requires. This is most evident in what are often referred to as "bloated government bureaucracies" (a fact that is partially attributable to the absence of the takeover option). But the same thing can and does happen in private corporations, though generally to a lesser degree.

Economic progress occurs most rapidly when there are strong pressures to produce the same output with less effort, to lay off workers when they are no longer needed. This often causes dislocations in the short-run, but in the long run it increases the availability of the most valuable resource (human effort and brainpower) to expand output elsewhere in the economy. So a strong case can be made that one of the advantages of the market for corporate control is that it increases the pressure on managers to keep the size of their workforce under control. If there were an active market for the control of government bureaucracies, where bureaucracy raiders could profit from the savings realized by eliminating redundant government jobs, does anyone doubt that these agencies would be run more efficiently – with far fewer workers?

Some of the efficiencies derived from hostile takeovers (and therefore some of the benefits to corporate shareholders) are the result of workers losing their jobs. But what is the extent of this loss, and which workers are most likely to be laid off? To address this question, 62 hostile takeover attempts (50 of which were successful) from 1984-1986 were examined.[22] According to this study, layoffs were common, but seldom exceeded 10 percent of the workforce, and were typically far less than that. Also, it was estimated that the probability of being laid off was 70 percent higher for white-collar workers than for blue-collar workers. The jobs of managers, not those of workers on the line, were

---

[21] Others have explained the advantages of moving toward more smaller and more focused firms in terms of improved, more efficient capital markets that have made it attractive for firms to substitute reliance on external capital markets for internal capital markets, which favor multi-division firms. See Amar Bhide, "Reversing Corporate Diversification," Journal of Applied Corporate Finance, Summer 1990, vol. 3 (Summer 1990), pp. 70-81.

[22] See Sanjai Bhagat, Andrei Shleifer, and Robert W. Vishny, "Hostile Takeovers in the 1980s: The Return to Corporate Specialization," pp. 1-72 in Martin N. Bailey and Clifford Winston (eds.) Brookings Papers on Economic Activity (Washington, DC: Brookings Institution; 1990).

most at risk. In addition, layoffs at targeted firms that were not taken over were greater (as a percentage of the workforce) than those in firms that were taken over. This suggests that the threat of a takeover provides a strong incentive for efficiencies even when no takeover actually occurs.

Takeover Defenses

Even if it is accepted that hostile takeovers are generally efficient, it doesn't follow that there should be no corporate defenses against such takeovers. Ideally there should be some resistance to takeover offers, but not "too much." Neither efficiency nor the interest of stockholders would be enhanced if the managers of a corporation simply acquiesced to the first takeover bid that offered more for the corporation's stock than the current price. The first bidder is not necessarily the one best able to improve the performance of the target corporation, and therefore the first bidder is not necessarily the one who can make the best offer. By being able to mount some defense against hostile offers, corporate managers can stimulate an aggressive auction that results in a winning bid that more accurately reflects the value of the corporation.

On the other hand, efficiency and the interests of shareholders can be harmed if the defenses against takeover bids are too impenetrable. If a takeover looks impossible, no one will make the effort to acquire control of even the most poorly managed corporation. Also, a significant investment is involved on the part of an outsider to determine the potential for improving the management of a target corporation and the maximum price that can be paid for its stock and still make the takeover pay. There is little motivation to incur the cost of this investment unless it gives those who do so a bidding advantage. So takeover defenses that go "too far" in requiring the initial bidder to make his information generally available can discourage takeover efforts to the point of reducing the amount of the winning bid.

No one can know exactly what is the best defense against a hostile takeover from the perspective of efficiency. Obviously the most efficient defense will vary from situation to situation. But some types of defenses that managers can mount seem to be more efficient than others.

Interestingly, there is evidence that bringing litigation against bidders increases the amount that is ultimately paid for the stock of the target corporation, assuming that the target corporation loses the case.[23] Managers of the target corporation can also defend against a takeover by offering to repurchase the stock acquired by a raider at a premiums; a practice known as *greenmail*. Some studies indicate that greenmail imposes significant negative returns on shareholders of the target (repurchasing) firm, but other studies indicate that greenmail can result in small gains for the repurchasing firm's shareholders.[24] Managers of the target corporation will want to be careful, however, if

---

[23]Recall, unless otherwise indicated the studies cited are discussed in Jarrell, Brickley, and Netter, "The Market for Corporate Control: The Empirical Evidence Since 1980."

[24]Michael C. Jensen and Richard S. Ruback, "The Market for Corporate Control: The Scientific Evidence," Journal of Financial Economics, vol. 11 (1983), pp. 5-50; and Wayne H. Mikkelson and Richard S.

considering a policy of greenmail, since any gain to shareholders probably comes by encouraging others to attempt a takeover in the hope of extracting greenmail. Paying greenmail on a consistent basis is obviously not a way of promoting the long-run profitability of a firm.

A very effective way for managers of a corporation to defend against a takeover is through what is referred to as *poison pills*. A poison pill describes a rule that allows shareholders of the target corporation to acquire additional shares at attractive prices, which serves to dilute the stock holding of the acquiring corporation. Although there are different types of poison pills, studies indicate that they are in general harmful to the wealth of the target corporation's shareholders.[25]

Managers can also protect themselves against takeovers by lobbying for legislation that reduces the chances that a takeover will be successful. Such legislation imposes a variety of regulations on takeover activity, but the studies that have been done suggest that, in general, they reduce shareholder wealth. The stock price of firms typically declines relative to the general stock prices when the state in which they are incorporated passes anti-takeover legislation.[26]

Obviously, the interests of managers and those of shareholders are not in perfect alignment in the case of takeovers. But there are possibilities for overlap that are worth noting. A justification for a controversial severance-pay contract for top managers is based on the desirability of reducing management opposition to takeover bids that benefit shareholders. Top corporate managers are commonly granted what are referred to as *golden parachutes*, which provide them with handsome compensation when they leave the corporation. Such compensation can be particularly useful in cases where top managers have to invest heavily in knowledge that is highly specific to the corporation, and therefore worth little elsewhere. Golden parachutes can also encourage executives to take greater risks, given that they know that they will receive a significant severance pay package if the risks they take result in losses and they lose their jobs.[27] The argument is that when these managers are offered generous severance pay they are less likely to oppose a takeover offer that promotes efficiency and increases shareholder wealth. Golden parachutes help bring the interests of top managers more in line with those of their shareholders. But as with all incentives, care has to be exercised. Golden

---

Ruback, "An Empirical Analysis of the Interfirm Equity Investment Process," Journal of Financial Economics, vol. 14 (1985), pp. 523-553.

[25] Paul H. Malatesta and Ralph A. Walkling, "Poison Pill Securities: Stockholder Wealth, Profitability, and Ownership Structure," Journal of Financial Economics, Journal of Finance, vol. 20 (1988), pp. 347-376.

[26] Michael Ryngaert and Jeffry Netter, "Shareholder Wealth Effects of the Ohio Antitakeover Law," Journal of Law, Economics, and Organization, vol. 4 (1988), pp. 373-383.

[27] In the absence of some form of handsome severance pay package, managers may be inclined to take too little risk, or less risk than the stockholders may want them to take. The stockholders can have diversified portfolios of stocks and companies over which they can spread their risks. Managers, on the other hand, can have a fairly narrowly invested portfolio, given that their talent, one of their biggest investments, is typically invested in one firm. Without some incentive to do otherwise, managers may be inclined to protect their investments by investing their firm's assets in safe ventures.

parachutes should not be so lucrative that they make an executive indifferent about keeping his or her job and losing it.[28]

Like all arrangements, golden parachutes can be poorly designed and abused. It may make sense to provide golden parachutes to no more than just the CEO of a corporation and a few members of the top-level management team. Typically, a significant number of managers are involved in facilitating a smooth transfer of control. But there is no reason to extend golden parachutes to managers not involved in such a transfer. Also, while golden parachutes can be too stingy to promote the shareholder interests, they can also be too generous from the shareholders' perspective. Ideally, golden parachutes will be provided only to those managers whose responsibilities are relevant to a takeover, and the severance compensation provided will be tied to premiums in share prices generated by the takeover.

There is at least tentative support for the proposition that golden parachutes promote the interests of shareholders. According to one study of corporations that adopted golden parachutes, corporate stock increases an average of about 3 percent when the adoption is announced. One interpretation of this result is that the golden parachutes increased the connection between the interests of shareholders and managers. It is possible, of course, that part of the increased stock value resulted from the belief that the announcement indicated that management was expecting a takeover bid and wanted to protect themselves against it.

* * * * *

The primary point of this chapter is that many so-called "hostile" takeovers are not really hostile, at least not from the perspective of the owners of the corporation being taken over. Throughout the chapter, we have suggested that hostile takeovers promote efficiency by encouraging managers to behave as good agents for their stockholders.

The efficiency of hostile takeovers will surely remain subject to debate. And certainly no serious person would argue that all hostile (or even friendly) takeovers are efficient. Mistakes are made in the market for corporate control that, after the fact, leave all parties worse off. So the debate over hostile takeovers will continue, and so will hostile takeovers. Of course, from the perspective of most managers, the fact that hostile takeovers will continue is more important than the debate over their efficiency. But the best way for managers to protect themselves against unwelcome attention in the takeover market is to do a good job enhancing the long-term profitability of the firm. And this is probably the best argument in support of the efficiency of hostile takeovers. Even if every hostile takeover that is attempted was itself inefficient, the fact that they can and do occur creates a strong incentive for managers to manage firms efficiently on behalf of their shareholders.

[28]For a more detailed discussion of golden parachutes, see Jensen, "The Market for Corporate Control."

Concluding Comments

Although the analysis of imperfect competition tells us something about the working of real-world markets, it does not answer all the questions economists have asked. The theories presented here have not done a good job of predicting the consequences of imperfect competition. Thus our conclusions regarding the pricing and production behavior of firms in monopolistically competitive and oligopolistic markets are tentative at best.

Economists seeking to make solid, empirically verifiable predictions about market behavior rely almost exclusively on supply-and-demand and monopoly models. Although predictions based on those models may sometimes be wrong, they tend to be easier to use and may be more reliable than predictions based on models of imperfect competition. Predictions aside, it is important to remember that most markets are imperfect.

In the Manager's Corner for this chapter, we tried to show how markets for goods can be affected by the market for capital. Indeed, the two markets are intrinsically bound up together. The competitiveness of the capital market – including the market for entire firms – will act as a discipline on managers who might believe that they can take advantage of their discretionary authority. Capital markets also induce managers to find the most cost-effective methods of production.

Review Questions

1. Under what circumstances could a monopolistic competitor earn an economic profit in the long run?

2. To achieve the efficiency of perfect competition, must a market consist of numerous producers? If not, what other conditions are required?

3. How does the number of producers in a market affect the chances of forming a workable cartel?

4. How do the costs of entering a market affect the chances of forming a workable cartel?

5. Must a monopolist employer share the monopoly profits with the managers and workers? If not, why not? If so, what does the "profit sharing" do to the monopolist's output level? Prices?

6. Should antitrust law attempt to eliminate all forms of imperfect competition? Why or why not?

7. "In an economy in which resources can move among industries with relative ease, a cartel attempting to maximize short-term profits will sow the seeds of its own destruction." Explain.

8. How would a cartel in a market for a network good collude on price? Explain.

9. Suppose that the managers of a firm allowed their internal departments to act as little monopolies or suppose that the managers paid their workers more than the labor market would bear. What would happen in capital markets? To the firm?

10. Would you expect government-run organizations to be more or less efficient than privately owned firms? Explain your answer with reference to capital markets.

CHAPTER 14

Business Regulation

If anyone can find such a thing as an "unregulated industry," he can sell it at a profit to the Smithsonian.

George Champion

Name an industry that has not, in some way, been under the authority of a government regulatory agency at some time. At the start of the century such a task would have been relatively simple. Today, with government extending its activities in all directions, it is not. Almost every economic activity either is or has been, at some time in the past, subject to some type of regulation at one stage in the manufacturing, wholesaling, or retailing functions. The list of federal regulatory agencies virtually spans the alphabet -- FAA, FDA, FEA, FPC, FRS, FTC, ICC, NTHSA, OSHA, SEC – to say nothing of the various state utilities commissions, licensing boards, health departments, and consumer protection agencies. As a result, it is much easier to list regulated industries than to name an unregulated one. Air transport, telephone service, trucking, natural gas, electricity, water and sewage systems, stock brokering, health care, taxi services, massage parlors, pharmacies, postal services, television and radio broadcasting, toy manufacturing, beauty shops, ocean transport, legal advice, slaughtering, medicine, embalming and funeral services, optometry, oyster fishing, banking, and insurance—all are regulated. Regulation was in the 1960s and 1970s, especially, one of the nation's largest growth industries (although there was something of a "recession" in regulations in the 1980s). Why have people been willing to substitute the visible foot of government for the invisible hand of competition?

Explaining regulation -- why and how it happens -- is a major challenge to economists.[1] Although several insightful theories have been proposed, statistical tests of those theories are incomplete and are at times based on crude data. Some instances of regulation or changes in regulatory policy cannot be explained by current theories. At best, we can only review what is known about regulation and project the economic results.

Today regulatory agencies are increasingly criticized by economists, businesspeople, consumers, and consumer advocates. The major concern is the extent to which regulation is designed to benefit the regulated industry. Some critics want more regulation, others less, depending largely on how they view the process of regulation.

---

[1] The major alternatives are reviewed in James C. Bonbright, Albert L. Danielsen, and David R. Kamerschen, *Principles of Public Utility Rates*, 2nd ed. (Arlington, VA: Public Utilities Reports, Inc., 1988), Ch. 2.

To understand the controversy surrounding regulatory policy, we must first understand the theory. This chapter begins with a brief description of several major federal regulatory agencies and then proceeds to the various theories.

Major Federal Regulatory Agencies

Federal regulatory agencies have existed for about a century. From their origins and functions we can learn much about the regulatory process. The four broad sectors of interstate commerce that have been regulated, in some cases for almost one hundred years, are communications, energy, transport, and urban services. Most regulating commissions—consisting of 3 to 7 members, typically appointed but sometimes elected—try to achieve basic economic goals of efficiency, and promoting certain social-political goals, including safety.

Beyond setting minimum and maximum prices, government regulations often control the entire rate structure of an industry. They may limit entry into the industry or stipulate what services and goods will be provided at what levels, and to whom. Regulatory approval is required to offer new services, or to expand, modify, curtail, or abandon a particular service. In short, regulation can -- and often does -- pervade all dimensions of production and distribution.

The Interstate Commerce Commission (ICC)

The Interstate Commerce Commission (ICC) was established n 1887 to deal with unfair business practices in the railroad industry. By the latter half of the nineteenth century, railroad companies had overbuilt and were engaging in cutthroat competition through customer rebates and price discrimination. In self-defense, several companies had formed a cartel to divide the market and set prices. The ICC was established to protect both consumers and small competitors and was supported by both the railroad and their customers.

Since then, the ICC's regulatory authority has been expanded to cover all motor carriers except airplanes engaged in interstate commerce—mainly trucks, boats, and buses. In the past, the commission has been authorized to set minimum and maximum rates. It is also responsible for ensuring adequate service. The seven members of the commission are nominated by the president and approved by the Senate for a term of seven years. No more than a simple majority of the commissioners may belong to the same political party, and a commissioner may be removed for "just cause," including conflict of interest.

Some muse that while regulation has tended to favor those who are regulated at the expense of consumers, even the regulated industries have been harmed by regulation. One economist put it this way:

> A good way to understand what has happened [to railroads] is to imagine a business that is prevented from adjusting its prices to changing market conditions and from negotiating with its customers. Furthermore, imagine that the business is not permitted to decide how much of its principal inputs to purchase, how much it will pay for them or even how to use them, and it may not decide where it will operate. Worse yet, imagine that it faces strong competitors who are not encumbered by similar constraints. It would be surprising if such a business survived at all. This is only a slight exaggeration of the railroads' position before 1980.[2]

For decades now, economists have advocated reducing the ICC's power. Finally, in 1980 the trucking and railroad industries were partially deregulated. Although the ICC no longer sets truck rates and routes, it still controls market entry through its authority to issue licenses.

The Federal Trade Commission (FTC)

The independent five-member Federal Trade Commission (FTC) was an agency established by Congress in 1914 to enforce the antitrust laws, especially the Clayton Act. The Antitrust Division of the Department of Justice is the other federal antitrust enforcement agency dealing especially with the Sherman Act. FTC commissioners are appointed and serve seven-year terms. To carry out their duties, they are given the power to probe through corporate records and summon corporate executives to hearings on unfair competitive practices. They can also issue formal complaints and order a company to cease its illegal acts. For example, state bar associations once restricted lawyers from advertising their services. The FTC ordered a halt to such restrictions on the grounds that they thwarted competition.

The Reagan administration tried to reduce the regulatory power of the FTC by cutting its budget—a ploy resisted by Congress. In the early 1980s, however, FTC decisions began to reflect the free market views of its new chairman, James C. Miller, a Reagan appointee who later served as the head of the Office of Management and Budgeting.

The Federal Communications Commission (FCC)

The Federal Communications Commission (FCC), established by the Communication Act of 1934, regulates telephone, telegraph, and broadcasting companies. Its seven commissioners, who are appointed for seven-year terms, set rates for interstate telephone and telegraph services and issue licenses to radio and television stations. The FCC determines who can engage in broadcasting, and it prescribes the nature of broadcast services, the location of radio and television stations, and the areas they serve. Licenses are issued for three years, after which the station's programming is reviewed for license renewal. To ensure renewal, a station must engage in some public-service broadcasting.

---

[2] "The Track Record," *Regulation* No. 1 (1987): 23—24.

To the extent that some available frequencies have not been put into use for radio and television transmission, the FCC has restricted entry into the broadcast business. It has also held up the introduction of cable service, which would vastly increase television programming variety. Yet in other ways the agency has sought to increase competition. At one time the American Telephone and Telegraph Company (AT&T) had a virtual monopoly over the sales of telephones. Beginning in the late 1960s, however, the FCC moved to introduce competition into the sale of telephone equipment and the delivery of long-distance service. In 1984, AT&T was separated from its twenty-two operating companies, which were consolidated into seven regional holding companies. AT&T maintained its manufacturing company, Western Electric, and the jointly owned Bell Laboratories. (See the Perspective on the AT&T break-up on page 21 in this chapter.)

The Federal Energy Regulatory Commission (FERC)

The Federal Power Act of 1930 established The Federal Energy Regulatory Commission (FERC). It is in the Department of Energy. Its authority was limited at first to the regulation of waterpower. In 1935, however, the FERC was authorized to regulate the rates, service, corporate practices, and security issues of interstate electric utilities. Beginning in 1938, it was empowered to fix rates for wholesale interstate natural gas service. At its zenith, FERC regulated electric, gas, gas and oil pipelines, and water power sites. The commission's five members serve five-year terms.

In the late 1960s and early 1970s, the FPC came under attack for its tight controls on the price of natural gas. In 1975, 1976, and 1977, several states experienced serious shortages of natural gas when the FPC-restricted price -- only one-quarter of the going price in producer states like Texas -- severely discouraged out-of-state sales. Natural gas was partially deregulated in early 1983, but was re-controlled in 1984 for two more years. Starting January 1, 1985, all gas discovered after April 20, 1977 was deregulated, while gas discovered before this date was—for the most part—not deregulated.

The Nuclear Regulatory Commission (NRC)

After the Atomic Energy Commission, which began in 1946, was abolished, The Nuclear Regulatory Commission (NRC) was established in 1974. The NRC licenses and regulates nuclear energy to protect the public health and safety, maintain national security, and comply with the antitrust laws. The NRC also sponsors a research program in reactor safety, fuel cycles, environmental protection, and so forth, and licenses imports and exports of nuclear materials.

The Securities and Exchange Commission (SEC)

In response to many instances of stock fraud, as well as the plunge in stock prices during the Great Depression, Congress established The Securities and Exchange Commission (SEC) in 1934. The SEC licenses stock exchanges and polices their activities. It has (but no longer exercises) the authority to regulate fees charged by brokers for carrying out their customers' transactions. In 1975, when the SEC decided to allow competitive

determination of stockbrokers' fees, those fees fell almost immediately by about 30 percent. The SEC also supervises the issuance of new securities by corporations and disclosure of information relating to those issuances. The commission has five members, who are appointed by the president for terms of five years. It has jurisdiction over securities and financial markets, and electric and gas utility registered holding companies.

The Food and Drug Administration (FDA)

Food and drugs have been regulated to some degree since the turn of the century. Not until 1931, however, was the Food and Drug Administration (FDA) established, as part of the Department of Health, Education, and Welfare (now called the Department of Health and Human Services). The FDA is responsible for ensuring the purity, safety, effectiveness, and accurate labeling of certain foods and drugs. No prescription or over-the-counter drug can be sold on the market before it has been judged safe and effective by the FDA. The agency is also responsible for enforcing a wide variety of consumer protection laws pertaining to the labeling, packaging, and advertising of foods and drugs.

The Occupational Safety and Health Administration (OSHA)

Probably no government regulatory agency is currently more controversial than the Occupational Safety and Health Administration (OSHA). Organized in 1969, in response to numerous reports that worker safety and health was not adequately protected, OSHA has formulated thousands of health and safety standards. To meet its requirements businesses have had to spend tens of billions of dollars. Those who believe government has an important role in protecting workers have praised OSHA, suggesting that if anything, the agency should conduct more inspections and impose higher fines to induce businesses to meet established standards. Businesses, on the other hand, have condemned OSHA's expensive standards as ineffective and wasteful.

The Public Interest Theory of Regulation

Regulation has often been justified on the grounds that it is in the public interest, meaning that it helps to achieve commonly acknowledged national goals. Some of the goals that may be pursued through regulation include:

- a more democratic allocation of the nation's resources (and a reduction in the importance of profit in such decisions);
- an increase in market efficiency;
- enhancement of the nation's ability to pursue certain essentially political objectives—improvement of the national defense, redistribution of costs of economic decisions, conservation of resources, and provision of certain public goods, such as public safety.

Economists' theories of regulation tend to be based on the goal of increasing market efficiency. One of the sources of market inefficiency economists cite most frequently is externalities, or third-party effects of market transactions.

Regulation to Capture Externalities

The market failure problems that externalities can cause were discussed much earlier. You will recall that an externality or spillover is a cost or benefit imposed on or enjoyed by other members of society by the activities of a producer or consumer that are not borne or enjoyed exclusively by the direct cause. An information disparity or asymmetry between producers and consumers is a form of market failure. Regulation is often imposed to ensure public safety, an economic good that is sometimes, but not always, an externality. Product features that ensure the safety of the purchaser—for instance, shock-absorbing steering columns—can be handled with reasonable efficiency by the market. Safety devices that benefit other persons, however, may not be provided by a market system.

For example, shock-absorbing bumpers benefit not only the person who buys a car but also those who may be involved in a collision with the buyer. If John collides with a car protected by shock-absorbing bumpers he may sustain less damage than he would have otherwise, without having paid for the protection received. He free rides on Mary's and the other driver's purchase. Because of the externality, the quantity of shock-absorbing bumpers purchased in an unregulated market will fall short of the economic optimum. Hence the need for regulation of safety equipment like shock-absorbing bumpers—and headlights, brakes, and/or windshield wipers.

Regulation sometimes benefits all producers, particularly when it enhances their reputation for safety. If people believe that a given product is safe, unscrupulous competitors may take advantage of the public's faith by reducing the safety of their products and cutting their production costs. Bad experiences with a product can make consumers skeptical of all firms, thereby reducing the price they are willing to pay for goods that may not prove to be safe. Thus by restoring consumer confidence, consumer protection laws can actually benefit the food and drug industries and toy manufacturers. To the extent that the SEC contributes to the securities industry's reputation for honesty, regulators can be seen as producers of public goods. However, externalities do not necessarily require government intervention. In certain cases a rearrangement of property rights may be more efficient.

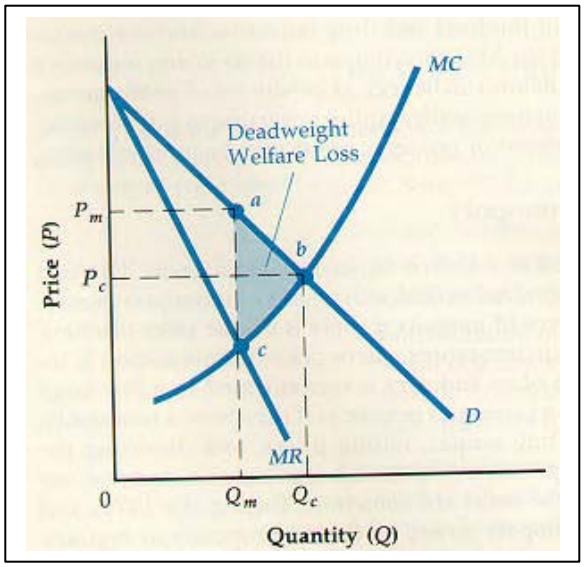
Regulation to Curb Monopoly

Monopoly is frequently cited as a source of market inefficiency. The first regulatory agencies were organized to deal with abuses of monopoly power. Monopoly can also be a source of inequity if there is undue price discrimination, although there are circumstances where price discrimination is socially optimal. If ownership of an industry is concentrated in a few large corporations, they can form a cartel and behave as if they were a monopoly, dividing the market, restricting output, raising prices, and distorting the price structure. To do this profitably, however, requires that demand initially be inelastic and entry be restricted somehow. During the 1970s and 1980s, the fear of such monopoly power motivated proposals to regulate the oil and automobile industries, among others.

Figure 14.1 shows a cartelized industry producing at an output level of Q_m and selling at a price of P_m. That output level is inefficient in two respects. First, it is less than the maximum, Q_c. Second, the marginal benefit of the last unit produced (equal to its price) is greater than its marginal cost. Although consumers are willing to pay more than the cost of producing additional units, they are not given the chance to buy those units. The cartel's price-quantity combination not only creates economic profit for the owners, which may be considered inequitable or unjust, but results in the loss of net benefits, or "deadweight welfare loss," equal to the shaded triangular area *abc*.

FIGURE 14.1 The Effect of Regulation on a Cartelized Industry

The profit-maximizing cartel will equilibrate at point a and produce only Q_m units and sell at a price of P_m. In the sense that consumers want Q_c units and are willing to pay more than the marginal cost of production for them, Q_m is an inefficient production level. Under pure competition the industry will produce at point *b*. Regulation can raise output and lower the price, ideally to Q_cP_c., thereby eliminating the deadweight welfare loss, equal to the triangle abc, resulting form monopolistic behavior.



Regulation can force firms to sell at lower prices and to produce and sell larger quantities. Ideally, firms can be made to product Q_c units and to sell them at price P_c, which is the same price-quantity combination that could be achieved under highly competitive conditions. At that output level, the marginal benefit of the last unit produced is equal to its marginal cost.

Government regulators need not demand that a company produce Q_c units. All they have to do is require it to charge no more than P_c. Once that order has been given, the portion of the demand curve above P_c, along with the accompanying segment of the marginal revenue curve, becomes irrelevant. The firm simply is not allowed to choose a price-quantity combination above point b on the demand curve. Then the profit-maximizing producer will choose to sell at P_c, the maximum legal price. With marginal revenue guaranteed at P_c, the firm will equate marginal revenue with marginal cost and produce at Q_c, the efficient output level.

Ideal results cannot be expected from the regulatory process, however. The cost of determining the ideal price-quantity combination can be extraordinarily high, if not prohibitive. Since regulators do not work for regulated industries, they will not know the details of a company's marginal cost or demand elasticity. The problem is particularly acute for regulators of monopolies, since there are no competitors from which alternative cost estimates can be obtained. Furthermore, if prices are adjusted upward to allow for a company's computed costs, a regulated firm may lose its incentive to control costs. To

the extent that regulators force prices below the level a regulated monopoly would otherwise charge, however, regulation serves the public interest by increasing market efficiency.

A call for regulation also has gone out to conserve scarce resources such as in radio and television broadcasting, natural gas, oil and water. Free market processes may result in overproduction relative to the perceived future societal needs.

The cost of the regulatory process must be emphasized. If regulation is truly to serve the public interest, it must increase the efficiency of the entire social system. That is, its benefits must exceed its costs. Too often the net benefits of regulation are overestimated because of a failure to consider its costs, which were estimated to exceed $100 billion in the early 1980s.

The Special Case of the Natural Monopoly

So far our discussion of monopoly power has assumed rising marginal costs (see Figure 14.1). One significant argument for regulation, however, is based on the opposite assumption. Some believe that in industries such as electric utilities, referred to as public utilities, the marginal cost of producing additional units actually decreases over the long run. That is, within the relevant range of the market demand, the long-run marginal cost curve slopes downward. **Subadditive costs** occur when a single firm can supply all the industry output demanded more efficiently than two or more firms can, making competition infeasible and creating a natural monopoly. In a **natural monopoly**, long-run marginal and average costs normally decline with increases in production, so that a single firm dominates production. Natural monopolies tend to be dominated by one firm, which will see monopoly profits once it is established as the sole producer. Natural monopolies are seen as prime candidates for regulation because their dominance in the market allows them to exert considerable monopoly power, provided demand is initially inelastic and entry is restricted. Table 14.1 shows the current status of the public utility sector, where the regulated firms were traditionally thought to be natural monopolies. As the table shows, though, this is not necessarily the situation today.

Assume, for example, that economies of scale lead to a long-run decline in the marginal cost of producing additional units of electricity. By producing on a larger scale, a firm can exploit the efficiencies of very large turbines to produce additional megawatts at a lower cost. Whether the size of generators can be increased indefinitely without producing diseconomies of scale is a matter of debate, as we will see later. Proponents of large electric plants believe that economies of scale are considerable—so extensive that in order to produce power at the lowest possible cost, only one extremely large electric company can operate in a specified geographical area. The fear is that once that firm emerges from the competitive struggle as the sole producer, it may be tempted to restrict production, charge a higher price, and reap monopoly profits.

The theory of natural monopoly is more fully explored below with appropriate graphs. Here we will simply note that it is unconvincing to many economists because it does not account for the presence of potential competitors. New firms, not currently competing in the market, may enter if the sole producer begins to extract economic profits through monopoly pricing. The possibility becomes more obvious if we think of a

natural monopoly as the only hardware store in a small town, or the only amusement park within several counties, rather than as a large electric company. While such producers are technically natural monopolies, they must fear the entry of competition enough to restrain their monopolistic tendencies.

TABLE 14.1 Traditional Public Utility Sectors and Their Current Status

Primary Monopolies	Primary, Party, or Potentially Competitive
Local telephone service	Long-distance telephone service
Local electric power distribution	Specialized postal services
Local natural gas distribution	Railroads
Basic postal services	Waterways
Cable television	Pipelines
Urban transit	Airlines
Water and sewage	Broadcasting
Ports	Hospitals
	Trucking

Source: William G. Shepherd, *Public Policies Toward Business* (Homewood, Ill.: Richard D. Irwin, 1985), Table 12-1, p. 330. Copyright © 1985 by Richard D. Irwin. Reprinted with permission.

As discussed in the previous chapter, a contestable market is a market—often multiproduct in nature—where ultrafree entry (and exit) constrains potential monopolistic behavior. [3] Contestability emphasizes market performance over market structure. Threatening credible potential entry (and exit) provides a weak "invisible hand" to induce efficient economic performance. The newer concept of contestability is similar to that of the older theory of workable competition in the sense of the analysis of the determinants of market performance. The major contribution of contestability may be in emphasizing the multi-product nature of modern businesses.

A Graphic Analysis of Natural Monopoly

To expand on our earlier discussion of the behavior of natural monopolies, we can use graphs to examine the arguments for and against regulation of this type of monopoly.

A Model of a Natural Monopoly[4]

[3] In a contestable firm, entry and exit is completely free; the costs and technology are the same for potential entrants as for existing incumbent firms; there are fixed but not sunk costs (unrecoverable from selling fixed inputs elsewhere); and buyers can purchase from the firm(s) that posts first the lowest price.

[4] Although today we know that economies of scale are neither a necessary nor a sufficient condition for a natural monopoly, we present this older approach as a tolerably accurate approximation to the more

As described earlier in the book, as the long-run marginal cost of production diminishes, the long-run average cost decreases as well, but at a slower rate. In Table 14.2, the marginal cost of producing each additional megawatt, shown in column 2, decreases from $50 for the first megawatt to $10 for the fifth. Though the average cost of the first unit is equal to its marginal cost, the average cost of subsequent units falls less rapidly than their marginal cost.[5] If we plot the marginal and average cost curves from the table on a graph, they will look like the curves in Figure 14.2.

TABLE 14.2 Long-run Marginal and Average Costs of Producing Electricity

Megawatts (1)	Long-Run Marginal Cost (2)	Long-Run Total Cost (3)	Long-Run Average Cost [(3) ÷ (1)] (4)
1	$50	$ 50	$50
2	40	90	45
3	30	120	40
4	20	140	35
5	10	150	30

Figure 14.3 shows these same curves along with the electric company's market demand and marginal revenue curves. According to traditional theory, a firm with decreasing costs will tend to expand production and lower its costs until it becomes large enough to influence price by its production decisions—that is, unit it achieves monopoly power. Then it will choose to produce where all monopolists produce, at the point where marginal cost equals marginal revenue. Thus the monopolistic firm in Figure 14.3 will sell Q_m megawatts at an average price of P_m, generating monopoly profits in the process. In other words, firms in decreasing-cost industries tend naturally toward monopoly.

Although a firm with decreasing costs can expand until it is the major producer in an industry, if not the only one, it will necessarily be able to manipulate price as a result. Suppose a natural monopoly flexes its market muscle and charges P_m for Q_m units. Another firm, seeing the first firm's economic profits, may enter the industry, expand production, and charge a lower price, luring away customers. To protect its interests, the firm that has been behaving like a monopoly will have to cut is price and expand production to lower its costs. It is difficult to say how far the price will fall and output will rise, but only one firm is likely to survive such a battle, selling to the entire market at a price that competitors cannot undercut. That price will be approximately P_1 in Figure 14.3.

rigorous notion of subadditivity of costs. When costs are subadditive, subsidies may not be necessary to get socially optimal results, but entry may need to be restricted.

[5] Remember, average cost is the total cost divided by the number of units produced. If the total cost of two megawatts is $90 ($50 for the first megawatt plus $40 for the second), the average cost of each megawatt is $45 ($90 divided by two units).

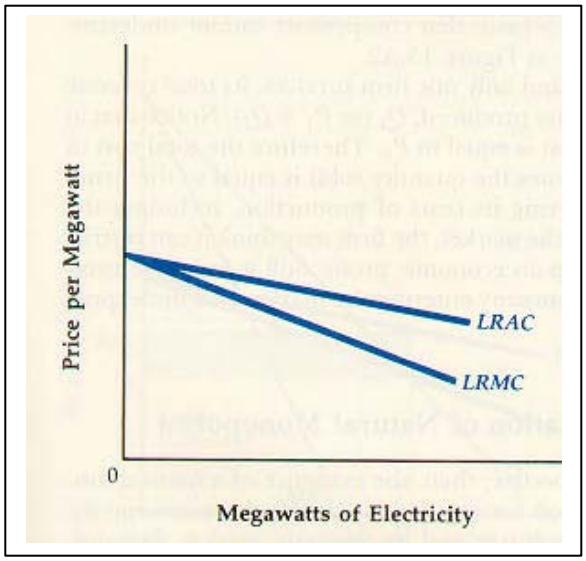


FIGURE 14.2 Long-Run Marginal and Average Costs in a Natural Monopoly

In a natural monopoly, long-run marginal cost and average costs decline continuously, over the relevant range of production, because of economies of scale. Although the long-run marginal and average cost curves may eventually turn upward because of diseconomies of scale, the firm's market is not large enough to support production in that cost range.

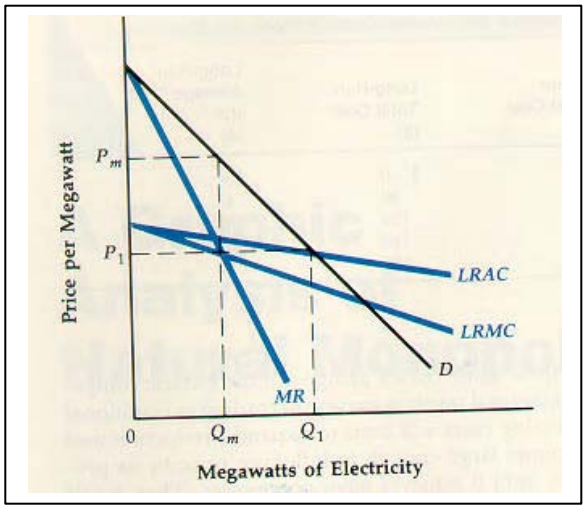


FIGURE 14.3 Creation of a Natural Monopoly

Even with declining marginal costs, the firm with monopoly power will produce where marginal cost equals marginal revenue, making Q_m units and charging a price of P_m. Unless barriers to entry exist, however, other firms may enter the market, causing the price to fall toward P_1 and the quantity produced to rise toward Q_1. At that price-quantity combination, only one firm can survive—but without barriers to entry, that firm cannot afford to charge monopoly prices. At a price of P_1, its total revenues just cover its total costs. Economic profit is zero.

If the price does fall to P_1 and only one firm survives, its total revenue will be its price times the quantity produced, Q_1 (or P_1 x Q_1). Notice that at that level, the firm's average cost is equal to P_1. Therefore the total cost of production (the average cost times the quantity sold) is equal to the firm's revenue. The firm is just covering its costs of production, including the owner's risk cost. Now alone in the market, the firm may think it can restrict output, raise its price, and reap an economic profit. Still it faces the ever-present threat of some other company entering the market and underpricing its product.

Arguments for the Regulation of Natural Monopolies

From a purely theoretical perspective, then, the existence of a natural monopoly is insufficient justification for regulation. Unless there are significant barriers to entry to an industry and an inelastic market demand, natural monopolies should not be able to charge monopoly prices. Proponents of regulation reply that some industries, like the electric

utilities require such huge amounts of capital that no competitor could be expected to enter the market to challenge the natural monopoly. That argument presumes, however, that the generation of electric power must take place on an extremely large scale. Such is not necessarily the case. Furthermore, the capital needed to produce electricity on a profitable scale can be raised by many large corporations, if economic profits exist.

Yet another argument for regulation—one more often voiced by the firms themselves than by consumers—is that production by natural monopolies generally requires large quantities of fixed capital assets, which become a sunk cost once purchased, to be ignored in short-run production and pricing decisions. If industries with long-run decreasing costs are vulnerable to destructive price wars, then firms that ignore massive fixed costs in the short run will eventually destroy themselves. This argument, however, presumes that entrepreneurs will enter an industry in which self-destruction is a likely outcome—a questionable presumption. In actuality, many industries—oil and automobile production, for instance—support a significant amount of competition despite extensive capital needs. Neither oil nor automobile producers seem likely to destroy themselves in the near future. However, modern transaction cost theory suggests that regulation may be needed as a contract arising because consumers need protection from monopolistic exploitation by producers, and producers need protection from opportunistic exploitation arising from the long-lived, transaction specific, idiosyncratic, immobile capital investments that are required to provide service.

Proponents of the regulation of natural monopolies point also to insufficient output and revenues. Even if an unregulated industry produces Q_1 units and prices that output at P_1 (see Figure 14.4), it has not reached the efficient output level. That would be the level at which marginal cost equals marginal benefit—the point at which the marginal cost curve intersects the demand curve. That level is Q_2 in Figure 14.4. Why does output fall short?

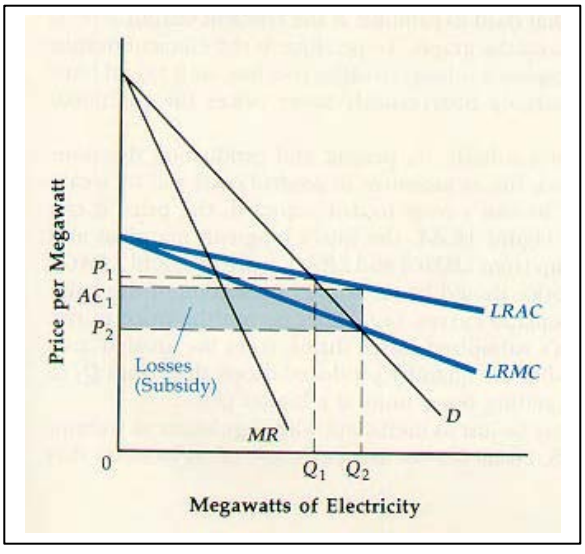


FIGURE 14.4 Underproduction by a Natural Monopoly

A natural monopolist that cannot price discriminate will produce only Q_1 megawatts, less than Q_2, the efficient output level, and will charge a price of P_1. If the firm tries to produce Q_2, it will make losses equal to the shaded area, for its price (P_2) will not cover its average cost (AC_1).

Given the market demand curve, the firm could sell an output of Q_2 for only P_2, earning total revenues of P_2 times Q_2. Since the average cost of producing at that output level -- AC_1 on the vertical axis -- would be greater than the price, total costs, at AC_1 x Q_2,would be greater than total revenues. The loss to a firm that tried to produce at the

efficient output level is shown by the shaded area on the graph. To produce at the efficient output level, a company would require a subsidy to offset that loss, or it would begin to price discriminate, charging progressively lower prices for additional units sold.

Once a firm is given a subsidy, its pricing and production decision must be closely monitored, for its incentive to control costs will be weakened. If the firm allows its cost curves to drift upward, the price it can charge will also rise. In Figure 14.5, the firm's long-run marginal and average cost curves shift up from $LRMC_1$ and $LRAC_1$ to $LRMC_2$ and $LRAC_2$. Following the rule that price should be set at the intersection of the long-run marginal cost and demand curves, regulators permit the price to rise from P_1 to P_2. The firm's subsidized losses shrink from the shaded area P_1abATC_1 to P_1cdATC_2—but the quantity produced drops also, from Q_1 to Q_2. Consumers are now getting fewer units at a higher price.

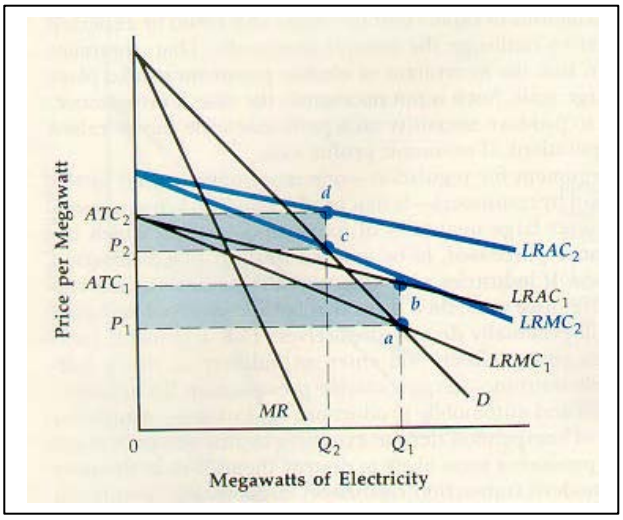


FIGURE 14.5 Regulation and Increasing Costs

If a natural monopoly is compensated for the losses it incurs in operating at the efficient output level (shaded area P_1ATC_1ba), it may monitor its costs less carefully. Its cost curves may shift up, from $LRMC_1$ to $LRMC_2$ and from $LRAC_1$ to $LRAC_2$. Regulators will then have to raise the price from P_1 to P_2, and production will fall from Q_1 to Q_2. The firm will still have to be subsidized (by an amount equal to shaded area P_2ATC_2dc), and the consumer will be paying more for less.

Thus, production may be just as inefficient with regulation as without it. Critics point to the U.S. Postal Service as an example of an industry that is closely regulated and subsidized, yet highly inefficient. Yet if the postal industry were truly a natural monopoly, it would be a low-cost producer and would not need protection from competition. Proponents of regulation see the inefficiencies we have just demonstrated as an argument for even more careful scrutiny of a regulated firm's cost—or for government control of production costs through nationalization.

Not all natural monopolies need subsidies to operate at an efficient output level. For all megawatts up to Q_1 in Figure 14.4, the unregulated firm can charge up to P_1, a price that just covers its costs on those units. If its product cannot be easily resold, the firm can price discriminate, charging slightly lower prices for the additional units beyond Q_1. As long as its marginal prices are on or below the demand curve and above the marginal cost curve, the firm will cover its costs while moving toward the efficient output level—and it can do so without giving other firms an incentive to move into its market. If its product can be resold, however, some people will buy at the lower marginal prices and resell to those who are paying P_1, cutting off the firm's profits.

Regulation of Destructive Competition

Another argument for government regulation is based on the existence of destructive, ruinous, or cutthroat competition. In direct contrast to the natural monopoly situation, where there is a shortage of competition, the destructive competition argument centers on a surplus of competition. In industries specialized as to location or purpose where there are high sunk costs in assets coupled with low operating costs, short-run bouts of intensive and perhaps destructive price cutting may emerge. Presumably, excess capacity triggered cutthroat competition in the early days of railroads. Competition among electric utilities, who must transmit power through wires, could mean several sets of power lines running down city streets, creating an environmental mess. This may be an argument for a government protected and regulated monopoly on the transmission of electricity, but it has no bearing on the need for regulation of the generation of electricity. The interstate pipelines for gasoline products are regulated, but there is competition among producers and refiners. Even if there were only one refiner, it would have to base its pricing decisions on what other firms might do if it tried to extract monopoly profits. The generation of electric power can be organized in a similar way. Duke Power, which serves parts of North and South Carolina, has proposed such a reorganization. As one of the nation's most efficient producers of electricity, Duke stands to expand its market share under a competitive system.

Regulation of prices is sometimes advocated as a safety measure. Some firms—for example, airlines and nuclear power companies—under competitive pressure to control costs may cut corners on safety. Regulation that keeps prices above competitive levels can induce such firms to compete in other ways—in terms of food quality, size of seats, or flight safety, for instance. Thus regulation can be seen as a means of correcting an under-production of safety. (If this argument is correct, the deregulation of airline rates in 1978 should have lowered the airline safety ratings.)

Critics of this theory suggest that a desire to avoid higher insurance premiums gives unregulated firms an incentive to maintain their safety precautions. Safety costs may not be completely internalized by insurance premiums, however, as illustrated by the 1984 accident at the Union Carbide plant in Bhopal, India, which killed over twenty-five hundred people and injured thousands of others. Given continued population growth and industrial concentration, regulation in the interest of public safety may be expected to increase.

The Evidence on Regulation

The public interest theory is not applicable to all forms of government regulation. Clearly environmental, traffic, and other safety rules promote public goals. Nevertheless, economists worry that such regulations can be used to thwart competition. Most environmental laws impose more stringent pollution standards on new sources of pollution than on old ones. The ostensible reason for this double standard is that new sources can meet the requirements at lower cost than old ones can, but such provisions can also be used as barriers to entry into competition.

Research has raised especially serious doubt about the usefulness to the public of economic regulation—regulation designed to restrict entry, pricing, and production decisions in specific industries like trucking, airline, bus, stockbrokerage, taxi, cable, and shipping services that appear to be competitive or contestable. In such industries, which were heavily regulated in the past, neither the prices charged nor the difficulty of entry can be justified on the grounds of efficiency. Much research, for example, suggests that the regulation of electric power companies has tended to prop up electric rates and to favor industrial and commercial users over residential users[6] Even in areas such as legal services and drugs, where the need for regulation has seldom been challenged, its value to the consumer is now being questioned.

Does regulation affect the competitive performance of an industry? The evidence is mixed, and open to differing interpretations. Many health, environmental, and safety regulations have clearly imposed substantial costs on businesses, consumers, and workers. Both the profits and the competitiveness of U.S. steel firms and the wages of steelworkers appear to have been seriously damaged by environmental legislation, for instance.

In the late 1960s and 1970s, regulation may have depressed the returns earned by electric utilities. Regulated industries have always had to wait for an upward adjustment of rates after a rise in costs. Apparently the unusually high rates of inflation during that period increased the strain on regulated industries. The story was different in the airline industry, however. As one researcher wrote, "Paradoxically the [Civil Aeronautics Board's policies, on the whole, have probably had little effect on the rate of profit earned by the industry; but, without the Civil Aeronautic Act and the Board, these profits would have resulted from quite a different sort of operation."[7] It was such arguments that led to the deregulation of the airline industry in the late 1970s. Other scholars have complained that FCC restrictions on entry into the broadcasting industry have enabled established broadcasting firms to make substantial profits.

In the trucking industry, regulation had particularly poor results. Until the industry was partially deregulated in the 1970s, the ICC turned down hundreds of applications a year to enter the trucking business or extend existing service. In fact, from the late 1930s through the 1960s, the number of licensed carriers actually decreased because of regulation by the ICC. Regulations designed to ensure a "stable trucking industry" frequently took trucks miles out of their way, increasing the cost of hauling cargo and the rates charged. After taking a load to one destination, carriers were forbidden to pick up cargo for the return trip.

The railroads had an entire century of regulation-induced problems. The results since deregulation in 1980 have been staggering. Prices have fallen, service has improved, profits have increased, and federal subsidies have fallen almost 90 percent.

---

[6] For reviews of empirical studies and conceptual arguments, see Paul W. MacAvoy, ed., *The Crisis of the Regulatory Commissions: An Introduction to a Current Issue of Public Policy* (New York: W.W. Norton, 1970); James Miller III and Bruce Yandle, eds., *Benefit/Cost Analysis of Social Regulation* (Washington, D.C.: American Enterprise Institute, 1979); and George C. Eads and Michael Fix, eds., *The Reagan Regulatory Strategy: An Assessment* (Washington, D.C.: Urban Institute, 1984)

[7] Richard W. Caves, "Performance, Structure, and the Goals of Civil Aeronautics Board Regulation," in *The Crisis of the Regulatory Commissions*, p. 134.

Additional gains will be more difficult as there are still many regulations in railroads, especially in labor-management relations.

Overall, the weight of the evidence is against much economic regulation. It is true regulatory agencies have sometimes denied rate increases and required firms—railroads and airlines, for example—to maintain services they would otherwise have eliminated. Many economists, however, question whether regulatory agencies as a group have been pursuing the public interest in any systematic way.

The Economic or Private Theory of Regulation

Why has regulation so often had little (if any) effect in reducing the profitability of regulated industries? Perhaps regulators have been inept at carrying out their responsibilities—or regulation may be too difficult a task for any one agency to handle properly. Regulated firms have an incentive to deceive their regulators by fudging their books to inflate their costs. As we have seen, gathering accurate information on a company's true costs and profits can be prohibitively expensive. Even with accurate accounting, there is an incentive to "gold plate" costs, as firms operate on a cost-plus basis. If demand is inelastic, these inflated costs can be passed on successfully to consumers. Moreover, regulation focuses on static efficiency and provides inadequate incentives for dynamic efficiency. A cost-saving innovation could lead to a cut in the utility's price.

A second explanation might be that while regulators are concentrating on prices and barriers to entry, firms may maintain profits by reducing the quality (and therefore the cost) of their products and services.

The intent of regulation may also be circumvented in another, more subtle way. Regulators sometimes determine prices on the basis of a so-called fair rate of return or profitability on capital investment. Such a standard encourages firms, particularly utilities, to substitute plant and equipment for other resources, such as labor, which do not count as investment. For example, suppose a regulatory agency establishes that 10 percent is a fair return on investment. Firms will then be allowed to make profits equal to 10 percent of the value of their plant and equipment. Suppose further that the same amount of additional electricity can be generated by spending $1 million on plant and equipment or $1 million on labor. If a firm invests in plant and equipment, it can ask the regulatory agency to raise its rates to allow for an additional $100,000 in profit (10 percent of $1 million). If it uses labor instead, it will have no increase in investment on which to base a request for a price increase. By making production capital-intensive, firms can circumvent the intent of regulation.

Thirdly, although regulation may be instituted with good intentions, regulators may become the pawns of regulated firms. If regulatory agencies are staffed by men and women who made their careers in the industries they are regulating, regulated firms may gain undue influence over regulatory policy.

Finally, the biggest shortcoming of regulation is that it often has been applied to competitive or contestable markets. Even if originally the market was a natural monopoly, it may have moved through a cycle where it is now competitive and thus no

longer in need of regulation. Many regulated industries are not now (and perhaps never have been) natural monopolies (e.g., motor trucking). In addition, some natural monopolies (e.g., main-frame computers) may have escaped the intricate web of regulation.

For all these reasons, many economists have begun to discard or at least downplay the public interest theory of regulation in favor of an industry-centered view. Instead of seeing regulation as something thrust on firms, they have begun to view it as a service frequently sought by those who are regulated.[8] It is important to recognize that the public and private interest theories are not necessarily diametrically opposed. The seeking of private interest is consistent with certain types of efficient regulation, and the public interest theory recognizes that mistakes and culpable regulators make regulation inefficient at times.

Probably the biggest impetus to the economic theory of regulation was the inadequacy of the public interest theory in answering two essential questions: Why were inherently competitive or contestable industries such as airlines, taxicab, and trucking regulated if the purpose was to protect against natural monopolistic pricing? Why do unregulated firms persistently desire to enter regulated industries if regulators push prices and profits to the bare-bones competitive level?

The Supply and Demand for Regulation

In the new expenditures theory of regulation, government is seen as a supplier of regulatory services to industry. Such services can include price fixing, restrictions on market entry, subsidies, and even suppression of substitute goods (or promotion of complementary goods). For example, regulation enables producers to suppress the sale of margarine in Wisconsin. Through the FCC, commercial television stations have been able to delay the introduction of cable TV.

These regulatory services are not free; they are offered to industries willing to pay for them. In the political world, the price of regulatory services may be campaign contributions or lucrative consulting jobs, or votes and volunteer work for political campaigns. Regulators and politicians allocate the benefits among all the various private interest groups so as to equate political support and opposition at the margin.

Firms demand regulation for their own private-interest, rent-seeking reasons. As we have seen, forming a cartel in a free market can be difficult both because new firms may enter the market and because colluders tend to cheat on cartel agreements. The cost of reaching and enforcing a collusive agreement can be so high that government regulation is attractive in comparison.

The view that certain forms of regulation emerge from the interaction of government suppliers and industry demanders seems to square with much historical evidence. As Richard Posner has observed,

---

[8] See George J. Stigler, "The Theory of Economic Regulation," in *The Citizen and the State* (Chicago: University of Chicago Press, 1975), and Stephen Breyer, *Regulation and Its Reform* (Cambridge, Mass.: Harvard University Press, 1982).

> The railroads supported the enactment of the first Interstate Commerce Act, which was designed to prevent railroads from price discrimination because discrimination was undermining the railroad's cartels. American Telephone and Telegraph pressed for state regulation of telephone service because it wanted to end competition among telephone companies. Truckers and airlines supported extension of common carrier regulation to their industries because they considered unregulated competition excessive.[9]

Barbers, beauticians, lawyers, and other specialists have all sought government licensing, which is a form of regulation. Farmers have backed moves to regulate the supply of the commodities they produce. Whenever deregulation is proposed, the industry in question almost always opposes the proposal.

Regulation as a Public Good for Industry

To the extent that regulation benefits all regulated firms, whether or not they contributed to the cost of procuring it, industries may consider regulation a public good. This creates a free-rider problem, which occurs when people can enjoy the benefits of a scarce good or service without paying directly for it by pretending not to want that good or service. Some firms will try to free ride on others' efforts to secure regulation. If all firms free ride, however, the collective benefits of regulation will be lost.

The free-rider phenomenon is particularly noticeable in large groups, whose cost of organizing for collective action can be substantial. Someone must bear the initial cost of organization. Yet because the benefits of organization are spread more or less evenly over the group, the party that initiates the organization may incur costs greater than the benefits it receives. Thus collective action may not be taken. Free riding may explain why some large groups, such as secretaries, have not yet secured government protection. Everyone may be waiting for everyone else to act. Small groups may have much greater success because of their proportionally smaller organizational costs and larger individual benefits. Perhaps it was because only a few railroad companies existed in the 1880s that they were able to lobby successfully for the formation of the ICC.

There are some exceptions to this rule. Several reasonably large groups, including truckers and farmers, have secured a high degree of government regulation, while many highly concentrated groups, such as the electrical appliance industry, have not. In highly concentrated industries. It may be less costly to develop private cartels than to organize to secure government regulation. In industries composed of many firms, on the other had, any one firm's cost of securing regulation may be smaller than the costs of a cartel. Large groups also control more sizable voting blocks than small groups. They may have the advantage of established trade associations, whose help can be enlisted in pushing for protective legislation.[10]

---

[9] Richard A. Posner, "Theories of Economic Regulation," *Bell Journal of Economics and Management Science* (Autumn 1974), p. 337.

[10] See Mancur Olson, *The Logic of Collective Action* (Cambridge, Mass.: Harvard University Press, 1971) Chs. 1 and 2.

In broad terms, the economic theory of regulation explains much above government policy—but that is one of its weaknesses. It is so broad as to limit its usefulness as a predictor. It does not enable economists to forecast which industries are likely to seek or achieve government regulation. Nor does it explain the current movement to deregulate the trucking and banking industries, or to regulate the environment. Neither of these trends appears to meet directly the demand of any particular business interest group. In general, any self-interested group will be better represented the larger its interest in the outcome, the smaller its size, the more homogenous its position and objectives, and the more certain the outcome.

Regulation as Taxation

According to a third theory, much of today's regulation can be explained as an indirect form of taxation—in the sense that taxation is the government's means of extracting money to pay for what are viewed as public goods and services. For example, until 1978, airlines were permitted to charge fares that exceeded their operations costs for long-haul flights. The extra revenues helped subsidize the below-cost pricing of short-haul flights and compensated airlines for their losses on unprofitable routes they were required to serve. In effect, some airline passengers were taxed to subsidize the fares of others.

In the postal service, another closely regulated industry, revenues from first-class postage have for years offset losses on magazines and bulk mail. Again, through regulation, one group of customers is taxed for the benefit of another. Seen this way, regulation appears to be a rather clumsy way of administering national tax policy—one that raises serious questions of equity in the distribution of the tax burden.

In general, transfers through "regulatory taxation" tend to go from dispersed to concentrated interests and are made as efficiently as possible, although inefficient transfers frequently occur. There is also a preference for disguising the costs imposed on victims of inefficient transfers and for broadcasting the benefits bestowed on recipients.

The Deregulation Movement

Recent years have seen a plethora of proposals to "deregulate"—actually "reregulate," as some type of government intervention still generally prevails—American industry. Airline (1978), trucking (1980), and railroad (1980) rates and routes have been deregulated. The price of natural gas was decontrolled in 1986, and the elaborate price controls on oil are more or less dismantled. Banks are now permitted to pay interest on checking accounts and are almost completely free to allow market forces to determine the interest rates on all their accounts. Surface freight forwarding had its entry, exit, and pricing deregulated in 1987.

Because economists have not extensively investigated the impetus for and results of deregulation, any assessment of the trend to deregulate must be considered tentative. In some cases, deregulation may have been a straightforward response to the inefficiencies of regulation. This seems a reasonable explanation for the deregulation of natural gas. The restricted supplies and shortages that characterized the industry under regulation were clearly not in the public interest.

The period of unusually rapid inflation in the late 1970s may also have encouraged the movement toward deregulation. In many industries, the process of seeking approval for price increases was cumbersome and time consuming, so that regulated prices lagged far behind the current rate of inflation. Under the circumstances, industry may have preferred the more competitive and flexible market system to the comparatively rigid regulatory system. This seems a reasonable explanation for the deregulation of truck rates and railroad routes in 1980. It may also explain why, beginning in 1980, banks were allowed to pay interest on checking accounts. Bankers may not have wanted to pay interest, but they had little choice, given the high returns depositors could earn on corporate and government bonds.

Another possibility is that regulated industries may simply have been outmaneuvered politically by consumer groups such as Common Cause and Ralph Nader's Public Interest Research Group. The votes of group members may have wielded more influence with Congress than industry's campaign contributions, especially after the size of political contributions was restricted. This may explain why the airline industry was deregulated in 1978 despite industry opposition. One regulatory agency, the Civil Aeronautics board (established in 1938), that had had economic control of commercial air transportation was even abolished in 1985.

It is possible, however, that regulation has not decreased overall. In the late 1970s, the visible foot of government was stepping into such new areas as the environment, worker health, and safety. These new regulations increased the effective tax on business, and thus the prices businesses charged consumers. Without doubt, the government's capacity to tax—that is, to impose costs on the private sector—is limited. Perhaps by deregulating some industries, the government reduced the effective tax in one area in order to increase it in others.

Economic theory suggests that whenever an industry is deregulated, there will be both gainers and losers. When the price of oil was decontrolled, for example, the losers were the consumers who found their purchasing power reduced by higher prices. Unless those who are hurt by deregulation are somehow compensated for their loss, they can create strong opposition to the change. One way the head off such opposition is to tax the gainers and subsidize the losers from deregulation. The windfall profits tax may be an example of such a scheme. When oil prices were deregulated in 1980, Congress imposed a heavy tax on profits it the domestic oil industry. The revenue from the tax was to be used for research on alternative energy sources like gasohol, and for low-income fuel subsidies.

Of course, the objectives and results of regulation cannot be evaluated solely in economic terms. Regulation may be intended to give citizens more influence on critically important decisions, such as the production of power, transportation, or defense readiness. Such objectives are essentially political, rather than economic, in nature.

PERSPECTIVE: ***The Break-Up of AT&T***

William F. Shughart, III, University of Mississippi

Before 1983, the U.S. telephone industry was a textbook example of a regulated natural monopoly. Once the basic switching equipment, trunk lines, and satellites are in place, the average cost of providing telephone service falls with increased output. Thus the industry came to be dominated by a single firm. Government regulation was justified as a way of controlling the monopolist's tendency to charge more than the marginal cost of service.

Although telephone service was regulated in the public interest, not all groups fared equally well under Federal Communications Commission (FCC) control. For example, the rate structure benefited local customers at the expense of long-distance customers. This cross-subsidy generally worked against commercial callers, whose demand for long-distance service was greatest during normal business hours, when rates were highest. Of course, AT&T benefited from barriers to the entry of new firms. But in the 1970s, the tables were turned, and AT&T itself became the victim of regulation.

Judge Harold Greene's historic decision ordering the breakup of AT&T followed a series of events that had been auguring change for over a decade. Most important was the development of microwave and satellite transmission technologies, which freed communications signals from earthbound telephone lines. In addition, since 1968 the FCC had been allowing customers to connect non-AT&T equipment to the Bell network. Throughout the 1970s it permitted new firms to compete with AT&T for long-distance service. AT&T was particularly hurt by the advent of competition in the long-distance market, which had long been among the most profitable of its operations. Discount carriers could charge less for the use of their long-distance transmission facilities mainly because they did not need to pay for switching equipment and local lines, which were owned by AT&T. In effect, MCI Communications Corporation and others were skimming the cream from AT&T's business.

By the late 1970s, then, the telephone industry was partly monopolistic (local service) and partly competitive (long-distance service)—and unworkable situation, from AT&T's perspective. One solution would have been to include the new carriers under the FCC's regulatory umbrella. The alternative was to break up Ma Bell, and this was the course advocated by the Department of Justice in its antitrust suit against AT&T, filed in 1974. In 1982 AT&T reached an agreement with the Department of Justice, approved by Judge Greene, which allowed it to retain its long-distance business. Its local business was divided among twenty-two local service companies. In return, AT&T was released from regulations that had prevented it from entering the computer business.

The history of AT&T shows clearly that regulation is not uniformly beneficial. Under deregulation increased competition has led to a proliferation of new telephone equipment and a decline in long-distance rates. Yet higher local rates and monthly access charges for long-distance service may wipe out those short-run gains.

Those who predict that local rates will eventually rise are assuming that before the breakup, AT&T was exploiting monopoly power only in the long-distance market. In other words, long-distance rates were set above marginal cost to make up for the revenue lost on local service. Differences in the profitability of the two markets may have stemmed from differences in the levels and elasticities of demand. If this latter view is correct, prices for local service may not rise.

The FCC apparently continues to view local telephone service as a natural monopoly. Local service companies retain the exclusive right to provide local service. They remain subject to regulation by a variety of federal, state, and local agencies. Yet increasingly, business customers have bypassed local companies by establishing their own in-house communication services. The fact that these arrangements are viable on a much smaller scale than that of a local telephone monopoly suggests that the natural monopoly argument may no longer be valid. In any case, the availability of alternative arrangements for telephone services will restrain the local monopoly's ability to raise prices.

In sum, the telephone industry is now in a period of transition characterized by rapid changes in both structure and technology, a phenomenon well into the 21st century. The future development of AT&T should provide some interesting examples of the effects of regulation and deregulation.

MANAGER'S CORNER: **The Value of "Mistreating" Customers**

Have you ever heard of a business consultant recommending to her clients that they mistreat their customers? Probably not. The standard recommendations consist of such advice as give customers what they want, pamper them, treat them as individuals, and never attempt to force them to do things they don't want to do. Most of the time this is surely sound advice. But not always. More often than not in business, consultants seem to realize, business can provide more value to their customers by mistreating them -- by giving them what they individually don't want, by ignoring their individual desires, by requiring that they do things they would not voluntarily do, and by charging them high prices for frills that cost more than they are worth.

If people always consumed services individually, with the value they received from their consumption unaffected by what others do, then mistreating them would seldom be a good business strategy. But many services are consumed either together, or in the presence of others. When this is the case, suppliers should always be alert to the possible *collective* benefits that can be realized by both them and their customers by mistreating them on an *individual* basis.

Putting Demands on Customers

In many cases, the benefit from mistreating customers is explained by the fact that by mistreating individual customers, a supplier allows the customers to overcome a prisoners' dilemma and be better off collectively. To see why, assume that you are the manager of a shopping mall that is soon to open for business and are anxious to attract retailers who will pay as much as possible for the opportunity to locate in your mall. This is a situation in which you should not be too accommodating to each potential customer, or tenant, in this case. A far better approach is one of creatively "mistreating" them -- requiring that they operate their stores in ways other than they would voluntarily choose if given a choice.

Hours of operation are one of the most important requirements you should impose on prospective tenants. It would be unusual if all tenants chose the same hours of operation. But you as manager would be smart to require that all tenants keep their stores open similar hours. The most obvious reason is there are significant costs involved in having the mall open, and it often doesn't make sense to incur those costs if only a few stores are open. You wouldn't want to keep a large mall open, for example, to accommodate a convenience store that wanted to stay open all night. This is why you don't find convenience stores operating in malls.

The most important reason, however, for requiring that all tenants in the mall operate similar hours is because it has the effect of lengthening the number of hours they are open. When one store is open for business, it attracts consumers that benefit other stores. Indeed, one of the primary reasons stores like to operate in malls is they each receive spillover business from customers who came to the mall to shop at other stores. But this means that when a store is open, it is creating benefits that it is not capturing entirely for itself, and therefore a benefit that it would ignore in its own decision to stay

open or close. This suggests that if left to decide on its own, each store would likely stay open fewer hours than is best from the point of view of all stores. As manager of the mall, it is your job not to ignore the spillover business that stores generate for each other. Every store can benefit if it is required to stay open longer hours than it would choose to on its own.

Consider a hypothetical example in which each store owner in the mall would independently choose to keep his or her store open 40 hours a week, with the result that each store earns profits of $1,000 per week. Assume also that if any one store increased its hours to more than 40 hours a week on its own, with all other stores staying with their 40 hour per week schedule, the store staying open longer would see its cost increase with very little additional business as a consequence. Its profits would fall to $900 per week. On the other hand, if all but one of the stores increased their hours to 48 hours per week, they would each increase their profits to $1090 per week, as the mall became more convenient for, and popular with, shoppers. But the one store that remained open only 40 hours would be able to free ride on the additional popularity of the mall and would then earn $1150 profit per week. On the other hand, if all stores operated 48 hours per week, all stores would earn $1100 profit each week. Total profits are greater if all stores stay open 48 hours (assuming there are more than 15 stores in the mall), but individually each store would choose to operate only 40 hours. As the manager of the mall, you will increase the value the mall provides tenants -- therefore the amount they are willing to pay in rent -- by going against the wishes of each tenant and imposing a 48-hour schedule of operation.

By imposing hours on all stores that are longer than any one would unilaterally choose, you have benefited all of the tenants by removing them from a prisoners' dilemma. A good mall manager will be constantly alert to other areas where he or she can require tenants to do things they would not individually choose to do (or prohibit activities they would individually choose to do), but which create a more profitable setting when done by all (or not done by any). For example, individual stores may profit from having clerks standing outside their stores' entrances and aggressively soliciting passing shoppers to come in. But if this became a common practice, all stores could suffer with consumers feeling less comfortable shopping at the mall and taking their business elsewhere. So all storeowners are collectively better off if all such solicitations are banned. They could earn more from a greater number of shoppers and more sales, so you could earn more in rent from the storeowners. On the other hand, a policy of requiring that each store in the mall advertise in the local paper (or on local TV and radio), more than any store would individually choose to do, can increase the profits of all by increasing the number of shoppers coming to the mall.

The situation at a mall is similar to that in a community of home owners who are subjected to a covenant imposing restrictions on such things as the color of the houses, the type and maintenance of the landscaping, and the number of cars that can be parked outside overnight. Almost everyone living in such communities dislikes some of the restrictions. Yet people are willing to pay more to live in communities with covenants because the cost to each family of abiding by the restrictions is less than the benefit realized from having the restrictions imposed on others.

Private schools face serious competition attracting customers. They have to cover their costs of educating students with tuition payments from parents who have the option of sending their children to public schools they have already paid for with taxes. Obviously private schools have to treat their customers well if they are to survive. But some of the most successful private schools recognize that treating their customers well as a group can require mistreating them individually. In many respects the education of children is a collective enterprise in which the best results require that all customers be required to do things that many would not voluntarily choose to do.

Consider the example of a private school in Nanuet, New York that has done very well in part because it has come up with a creative way of mistreating its customers. Love Christian Academy requires that all the parents have monthly meetings with their children's teachers and volunteer to work at the school at least one day a year. If parents miss, or are even late for, a meeting, they are fined $100. Parents who are fined, or who must attend a meeting on the night of their favorite TV programs to avoid the fine, often feel mistreated. One parent was quoted as "not pleased" with being fined for violating one of the rules, and some parents have removed their children from the school because of the strict rules. But the school thrives because most parents feel more than compensated by knowing that their children are attending school with other children whose parents are actively involved in their education.[11]

Similarly, few parents want their children spanked at school. But if the choice is between sending their children to a school where none of the students are spanked or to one in which any student who misbehaves is spanked, including their own, many parents prefer the latter. This is recognized by many private schools that advertise the fact that they believe in maintaining discipline in the classroom by subjecting unruly students to an old-fashioned spanking. Dr. Connie Sims, the superintendent of Love Christian Academy, makes clear that before students are accepted their parents must accept the school's disciplinary policy.[12]

While no one feels good about his or her children receiving poor grades, many prefer a school in which that possibility is likely to one in which it is unlikely. The school that holds its students to a high standard of academic achievement and only gives good grades to those who achieve that standard will have a better reputation than a school that doesn't. So while the students, and their parents, may feel mistreated if they receive poor grades, they prefer a school with a policy of giving low grades because of the additional educational value created by that policy.

Manufacturers who sell their products through independent dealers often impose restrictions on the price the dealers can charge for the products or the number of dealers who can sell them in a given area. These restrictions are referred to respectively as *resale price maintenance* agreements and *exclusive dealing* arrangements. The effect of these restrictions is to increase the price consumers pay, and for a long time the conventional

---

[11] See Steve Stecklow, "Evangelical Schools Reinvent Themselves by Stressing Academics" The Wall Street Journal, May 12, 1994: p. A1.

[12] Ibid. We want to emphasize that our concern here is not whether or not spanking is the best, or even a good, way of disciplining children. The point is that many schools can attract business with practices that each of their customers would find objectionable if applied only to their children, but which they appreciate when applied to all students.

view of policy critics was that the price maintenance agreements and exclusive dealerships allowed sellers to profit at the consumers' expense. But, as in the previous examples, a policy that at first glance appears to be mistreating customers may actually be in the customers' best interest by allowing them to overcome a prisoners' dilemma.

In certain cases, requiring retailers to charge higher prices (price maintenance) or allowing them to charge higher prices (exclusive territory) makes it possible for a manufacturer to benefit customers because without these restrictions each customer would find it individually rational to behave in ways that are collectively harmful. Consider a product on which customers are able to make a more informed choice when it is properly displayed. One example is furniture, which is best examined in a well-appointed setting containing other pieces of complementary furniture.

Another example is sound equipment that consumers would like to evaluate in sound rooms before purchasing. But without the manufacturer being able to impose some restrictions on the retailer, it is unlikely that the consumer will benefit from such helpful displays. The retailer who went to the expense of properly displaying a product or having experts on hand to answer questions of potential customers would be vulnerable to the price competition of retailers who did not provide these services. A retailer with a warehouse and an 800 number could (and many have) run advertisements suggesting that customers visit retailers with showrooms and experts to decide what they want to buy, and then call in their order at a discount price.

The problem is that while it makes sense for each customer to take advantage of such offers, if many customers do so they will end up collectively worse off as the retailers with showrooms go out of business. This is clearly an example of consumers finding themselves in a prisoners' dilemma. So retail price maintenance agreements and exclusive-dealing arrangements can be thought of as ways of protecting consumers against their own prisoners' dilemma temptations. By not selling their products through a retailer who refuses to maintain some minimum price, a manufacturer can prevent some retailers from free riding on the showrooms and expert sales staffs of others. If price competition is not permitted, retailers must compete through the display, service and sales expertise that make the product more valuable to consumers. Similarly, by providing one retailer the exclusive right to sell its product in a market area, a manufacturer prevents, or at least reduces the ability of, some retailers to free ride on that retailer's efforts. A retailer with the exclusive right to sell a product in an area has a strong motivation to provide the combination of display and service that consumers find most attractive. And with each consumer able to secure the advantages of good displays and service only by paying for them, they are no longer in a prisoners' dilemma.

There is no guarantee, of course, that a manufacturer will choose a price (in a resale price agreement) or a market area (in an exclusive dealing arrangement) that makes consumers better off than they would be without such restrictions on retailers. For example, the resale price agreement could require a price that cost the consumer far more than the extra sales and service is worth. Or the exclusive market area could be so large that many customers are inconvenienced by the lack of a nearby store carrying the product. But a manufacturer who makes such mistakes will find itself penalized by competitors who make better use of these restrictions on retailers. Those manufacturers who strike the best balance between "mistreating" their customers with higher prices and

restrictions on the number of retailers in protecting their customers against the collectively harmful temptations of the prisoners' dilemma will expand their market share at the expense of those who do not.

Manufacturer restrictions on retailers will not make sense for all products. And manufacturers should be aware that the use of these restrictions might activate over-zealous antitrust enforcers even when they do make sense. But such restrictions do provide another example of how you can attract more business with policies that may appear to harm customers, but which actually benefit them by helping them escape a prisoners' dilemma.

Why the Customer Is Not Always Right

One of the oldest sayings in business is "The customer is always right." This seems like good advice to a firm that wants to succeed in the market place. Even if you believe your customers are wrong, don't disagree with them. Give them what they want, or they will take their business to someone who will. There are situations, however, when the only way to succeed is by being willing to tell your customers that they are wrong, and to give them exactly what they don't want.

Consider the situation faced by firms that are in the business of rating the bonds of corporations. Corporations that want to issue bonds pay firms such as Standard and Poors, Moody's, Duff and Phelps, and Fitch to evaluate the safety of those bonds and rate them accordingly. A rating of AAA indicates that the bonds are very safe, while a rating of CC indicates that the bonds are in the category of junk bonds and highly risky. A corporation does not want to misrepresent the safety of its bonds since doing so would, in the long run, reduce its ability to borrow money. But there is a natural tendency for a corporation to give itself the benefit of the doubt and believe that its bonds are safer than they actually are. Therefore, the rating service that followed the advice, "The customer is always right," would seriously jeopardize its usefulness, and its profitability. The rating service that developed a reputation for yielding to client pressure for higher ratings would cease to have the credibility that its clients are paying for. So, in the bond rating business, corporations commonly give good money for bad ratings.

Similarly, corporations hire independent accounting firms to audit their financial statements and report on the degree to which those statements conform to acceptable accounting practices and accurately convey relevant financial information. The managers of the corporations who pay for an audit have objectives (rapid promotions, nicer perks, and higher salaries) that differ from those of the stockholders, bondholders, and others who use corporate financial statements (and who want the highest return on their investments). So managers can have an incentive to bias the financial statements in ways that make them look better but are misleading to investors. But the accounting firm that does the audit has a strong incentive to ignore any desire managers may have for a favorable but undeserved report by being as impartial and accurate in its evaluation as possible. Only by maintaining a reputation for impartiality and accuracy is an accounting firm able to provide a valuable service to all of its clients.

Raising Price to Increase Customer Appeal

One of the best-documented rules of business is that the lower the price charged for a product, the more of that product consumers will buy. Everything else equal, consumers do prefer low prices to high prices, and it would seem that intentionally charging higher prices for products than they are worth would be a better way of driving away customers than attracting their business. But everything is not always equal, and there are situations where a business is well advised to charge its customers high prices to cover the costs of products they don't value that highly.

The benefits a person receives from consuming a good or service are sometimes significantly influenced by whom the other consumers are. Consider a rather extreme example. There are two hotels in the town you are visiting that are identical except for their customers. One is patronized by non-affluent and poorly behaved rowdies who create loud disturbances all night, while the other is patronized by affluent, well-behaved folks who are careful not to disturb their neighbors. Which hotel would you prefer? Preferences differ, and no doubt some would prefer the action that is more likely available at the first hotel. But it is a safe bet that most affluent, well-behaved people would prefer and be willing to pay more for the second.

This situation suggests what looks like a profit opportunity for one of the hotel owners; establish a reputation for catering to the affluent and well-behaved guests by refusing to rent to anyone else, and then charge premium prices. Unfortunately, things aren't so simple. First, it is not easy to tell if a prospective guest is either affluent or well behaved, particularly those who make telephone reservations. Second, even if you could identify those who are "unacceptable," refusing to rent to them would probably be a violation of public-accommodation laws in your state.

But there is another way to filter out less desirable customers that, though imperfect, has the advantage of not being illegal and of getting immediately to your primary objective. Just charge higher prices than the other hotel, even though it is physically identical. The less desirable customers will tend to take their business to the other hotel, which makes your hotel more valuable to those who can afford to pay extra to avoid the less affluent and/or unruly guests. As indicated, this strategy won't work perfectly. It does not, for example, screen out rock bands that may be affluent but very unruly. But though imperfect, high prices do have the virtue of generally doing a good job of screening out less desirable guests, and this is clearly a case where virtue is its own reward.

Things are more difficult, however, than indicated so far. All hotel owners would like to increase their profits by simply increasing their prices and catering to the well to do. Obviously, not everyone can be successful with this strategy. Because of competition, those who want to attract the well to do to their hotels with higher prices will find that they also have to provide nicer facilities and more services than are available at lower-priced hotels. So construction and operating costs will increase at high-priced hotels until the return on investment in these hotels is about the same as the return on investment in low-priced hotels, as well as in most other investments. But because one of the big benefits to guests at expensive hotels is being in the company of

other guests who can afford to pay high rates, the frills at those hotels don't have to be worth what they cost.

Indeed, it is widely believed that people pay more for extras than they are objectively worth at expensive hotels. One of the cut-rate hotel chains recently took advantage of this belief in an advertisement in which a hotel guest is shown holding up a small bottle of fancy shampoo and asking whether it was worth the extra $20 room charge. If not, the listener was urged to stay at the cut-rate hotel rather than one of the expensive hotels. A clever advertisement, but it ignores the fact that people are getting more for the extra $20 than the shampoo. They are getting a place to stay that screens out those who aren't willing or able to pay an extra $20 for a small bottle of shampoo.[13]

We have confined our discussion to hotels so far, but there are other businesses where the client effect is important in determining how much value consumers realize from the service. The client effect is certainly important for many people when they go out for a leisurely dining experience. People pay a lot of money for a meal in a fine restaurant, and though the food and service is typically quite good, it seems reasonable to wonder if many people actually value the attention of hovering captains, wine stewards, and waiters as much as they pay for them. Surely some of the benefits customers receive from the high prices at fine restaurants come from the screening performed by those prices. The client effect is hardly a consideration when you grab your food in a paper bag at a drive-through window. At McDonalds or Burger King the price you pay reflects the value you place on the food, not the value you place on screening out undesirable customers.

The business of education is another example of the importance of the client. Students who attend a college with other students who are capable and enthusiastic will typically get a far better education than those who attend a college with students who are poorly prepared and uninterested, even though the colleges are of similar quality in terms of faculty and facilities. Students learn not only from their classroom experiences, but also from their after-class interaction with other students. This suggests that the high tuition charges at many small private colleges can be explained, at least in part, by the value they create as screening devices.

We should point out that the screening explanation for high tuition is one that we find attractive. Both of the authors have spent their careers teaching in public universities where the students pay relatively low tuition. We have often wondered why so many small private colleges could charge such high tuition when, generally, most of the professors at these colleges, at least in the academic fields with which we are familiar, have published less and are less well known than our colleagues. Why would students, or their parents, pay so much more to attend the lectures of these professors when they could be attending lectures at our universities for far less? We certainly don't want to believe

[13] We don't want to overemphasize the difference between the costs of providing a package of extras (or frills) at expensive hotels, and the value of those extras to guests. Because of competition, hotels are strongly motivated to provide those extras that, for any given cost, provide as much real value to their guests as possible. But this is consistent with a hotel being able to realize a competitive advantage by increasing the supply of extras into the range where the extras themselves are worth less to the guests than they are paying because of the screening benefit provided by the extra charge.

that our colleagues are not as good at teaching as professors at expensive private colleges. Generally speaking, our colleagues are good teachers.

In our more serious moments we recognize, of course, that private colleges typically put more emphasis on teaching than do large public colleges and universities. Competition drives private colleges to provide more services to their customers for the extra price they pay. But it is hard to believe that the value created by the extra emphasis on teaching at these private colleges is nearly enough to justify the extra tuition charged. Surely much of the extra value is created by varying the higher tuition as a screening device.

The Link between Customer Abuse and Worker Wages

When bosses repeat the refrain "The customer is always right," workers may be led to believe that the unspoken rule is that they should take whatever the customers throw at them in the way of abuse. As we have seen, the bosses' advice might be a reasonable working rule, but it is also likely to be advice that the boss doesn't want employees to take with complete seriousness. The rule overlooks the fact that abusive customers can make work a form of "hell" for the workers. If forced to take excessive abuse, the workers would, no doubt, demand higher wages to compensate them for this abuse. At some point, as more and more abuse is encountered, it is altogether reasonable to expect that the higher wages the workers require will exceed the value received by the firm from accommodating abusive customers. Any tolerably reasonable boss will, at some point, ask workers to stand their ground and return the "fire" of their customers. Otherwise, firm profits can be impaired.

The president of Southwest Airlines understands the (economic) principles at stake. He has been known to write letters to customers who have been abusive to his workers, telling those customers that they should take their business elsewhere. Southwest may lose some business, but they can also gain a total wage bill that will be lower than otherwise, and that can more than compensate the company for the lost business. Also, the policy may screen out unruly passengers, thus making Southwest more attractive to well-behaved passengers.

Indeed, if customers are given too much consideration, some will abuse it at the expense of not just the business, but of customers in general. Consider refund policies. Most retail stores allow customers to return merchandise that they feel doesn't suit their needs as well as they anticipated. Within reasonable limits, such policies benefit all customers and build goodwill and profitability for the business. Some retailers have pushed those limits, however, with almost no restrictions on refunds. Apparently, some retailers are now having second thoughts as more and more customers are taking advantage of generous refund policies.

For example, Best Buy has stopped giving refunds on certain products unless the customer has a sales receipt, and even then the customer has to pay a "restocking fee" of 15 percent of the purchase price if the package in which the product came has been

opened.[14] Before the change in policy, one Best Buy customer received a refund on a video recorder that he claimed was defective. Indeed, it was defective for a reason the Best Buy repair technicians discovered when they played back the tape inside and saw the splash of water as the camera fell into a swimming pool and sank to the bottom. It was at the bottom when the recording stopped. Wal-Mart has also moved away from its open-ended return policy by imposing on most items a 90-day maximum beyond which no refund will be made. Before this restriction went into effect, a customer got a refund for a beat-up thermos that Wal-Mart later learned from the manufacturer had been purchased in the 1950s, long before there was a Wal-Mart. Another retailer that has decided to halt its no-questions-asked policy on returns is the catalog store L.L. Bean, Inc. According to a spokeswoman for the firm, some customers were returning clothes that had been purchased at garage sales or found in the closets and attics of deceased relatives.[15]

Most customers are honest, and a largely unrestricted return policy would be appropriate for them. But honest people will be the most supportive and appreciative of restrictions when a liberal return policy begins to be abused. And there is a tendency for the number who take advantage of a generous return opportunity to grow over time as some of those who do not initially return items that shouldn't be returned see others doing so. The cost of paying people for fraudulent, or at least highly questionable, returns is soon reflected in the price that everyone has to pay. Imposing strict limits on all customer returns will seem like mistreatment to some, but it is really little different than imposing restrictions on the hours of stores in a mall or fines on parents who are late for meetings with their children's teachers. Without such restrictions, each consumer will have an opportunity to gain by engaging in behavior that is collectively harmful.

* * * * *

Treating customers as if they are always right, giving them what they want, and giving it to them at the lowest possible price is standard business advice, and it is generally sound advice. But not always. We have examined several situations in this chapter where, when compared with the standard advice, good business calls for "mistreating" customers. Of course, once the situations have been explained, the recommended treatment of customers isn't mistreatment at all. Business owners and managers are well advised to be constantly on the alert for creative ways of "mistreating" their customers. There are many more circumstances where such creative "mistreatment" can allow a business to better serve its customers than can be known by any one person or discussed in one chapter. "Mistreatment" often is in the customers' interest and translates into economic improvement for customers and a higher value for the firm.

Concluding Comments

Treating the public interest and (private) economic theories of regulation separately may have suggested that one or the other must be the correct theory of regulation. In the real

[14]Louis Lee, "Without a Receipt You May Get Stuck With That Ugly Scarf," Wall Street Journal, November 18, 1996, p. A-1.

[15] Ibid.

world, however, the sources of regulation are complex. For instance, a combination of forces probably motivated the regulation of the airline industry. Some people were pursuing the public interest, as they saw it. Others, especially those connected with the airlines, saw an opportunity to protect their markets. The precise nature of any particular regulation probably reflects the relative strengths of these two forces, as well as the extent to which government allows them to be expressed. Moreover, these various theories are not diametrically opposed: For example, both public interest and the private interests may, at times, willy-nilly, promote efficiency or inefficiency. Neither approach has a monopoly on the truth. Each explanation mirrors a facet of reality. Neither one is valid standing alone.

Probably the most important lesson from the study of regulation is that while the public interest -- especially as it relates to the improvement of market efficiency—is a valid basis for regulation, it can be easily exploited. Pretending to pursue the public interest, promoters of regulation can realize their own interest instead. The statistical studies cited in this chapter indicate a considerable tendency to abuse the intent of regulation. How than can government serve the public interest without allowing a great deal of freedom for the special-interest regulation to have the government do their bidding? How can government regulate the regulators, and do so efficiently? These are the questions that must be addressed by any movement for regulatory reform.

Review Questions

1. The economic theory of regulation suggests that firms have an incentive to support protective regulation. Do workers have a similar incentive? Under what conditions would they support protective regulation, and under what conditions would they oppose it?
2. Aircraft producers once supported government efforts to hold airfares above competitive levels. Explain their position. Should the fare restrictions have led to more profitable airlines in the long run?
3. Develop a public interest case for the regulation of barbers and beauticians.
4. If a regulatory agency determines electric prices on the basis of a fair rate of return on investment," how might its price-setting standard affect the use of fuels in producing electricity? Would fuel oil producers favor the fair-rate-of-return method for regulating electric utilities?
5. In the 1970s, regulatory agencies allowed electric companies to pass on to customers any increases in the cost of their fuel -- a scheme that reduced companies' incentives to reduce fuel consumption and costs. Considered in the overall context of regulation, however, are such fuel adjustments necessarily inefficient? What might be the alternative to automatic increases based on the cost of fuel?
6. Economists argue that if utilities charge higher prices for electricity during period of peak demand, consumers will use less electricity then, reducing the strain on generators. Assume again that electric rates are based on the fair-return method of

regulation. Will electric utility companies favor the institution of peak-load pricing? Why or why not?

7. From the data in the following table, plot this natural monopoly's marginal cost, average cost, and demand curves on a graph. Label the efficient output level. Will the firm actually produce at that level? Why or why not?

Quantity	Marginal Cost	Price Consumers Will Pay
1	$21	$45
2	18	36
3	15	27
4	12	18
5	9	9
6	6	0

CHAPTER 15

Competitive and Monopsonistic Labor Markets

Labour, like all other things which are purchased and sold, and which may be increased or diminished in quantity, has its…market price

David Ricardo

Professional football players earn more than ministers or nurses. Social workers with college degrees generally earn less than truck drivers, who may not have completed high school. Professors of accounting typically earn more than professors of history with equivalent educational background and teaching experience. Even if your history professor is an outstanding teacher, capable of communicating effectively and concerned about students' problems, she probably earns less than a mediocre teacher of accounting.

Why do different occupations offer different salaries? Obviously not because of their relative worth to us as individuals. Just as there is a market for final goods and services—calculators, automobiles, dry cleaning—there is a market for labor as a resource in the production process. In this competitive labor market, the forces of supply and demand determine the wage rate workers receive.

By concentrating on the economic determinants of employment—those that relate most directly to production and promotion of a product—we do not mean to suggest that other factors are unimportant. Many noneconomic forces influence who is employed at what wage, including social status, appearance, sex, race, and personal acquaintances. Our purpose is simply to show how economic forces affect the wages paid and the number of employees hired. Such a model can show not only how labor markets work, but how attempts to legislate wages, like minimum wage laws, affect the labor market.

The general principles that govern the labor market also apply to the markets for other resources, principally land and capital. The use of land and capital has a price, called rent or interest, which is determined by supply and demand. Furthermore, land, capital, and labor are all subject to the law of diminishing marginal returns. Beyond a certain point and given a fixed quantity of at least one resource, more land, labor, or capital will produce less and less additional output.

The Demand for and Supply of Labor

Labor is a special kind of commodity, one in which people have a personal stake. The employer buys this commodity at a price: the wage rate the laborer receives in exchange

for his or her efforts. In a competitive market, the price, or wage rate, of labor is determined just as other prices are, by the interaction of supply and demand. To understand why a person earns what he does, then, we must first consider the determinants of the demand and supply of labor.

The Demand for Labor

The **demand for labor** is the assumed inverse relationship between the real wage rate and the quantity of labor employed during a given period, everything else held constant. The demand curve for labor generally slopes downward. At higher wage rates, employers will hire fewer workers than at lower wage rates.

The demand for labor is derived partly from the demand for the product produced. If there were no demand for mousetraps, there would be no need—no demand—for mousetrap makers. This general principle applies to all kinds of labor in an open market. Plumbers, textile workers, and writers can earn a living because there is a demand for the products and services they offer. The greater the demand for the products and the greater the demand for the labor needed to produce it -- and the greater the demand for a given kind of labor, everything else held equal, the higher the wage rate.

The productivity of labor -- that is, the quantity of work a laborer can produce in a given unit of time—is another critically important determinant of the demand for labor. The price of the final product puts a value on a laborer's output, but her productivity determines how much she can produce. Together, labor productivity and the market price of what is produced determine the market value of labor to employers, and ultimately the employers' demand for labor.

We can predict that the demand for labor will rise and fall with increases and decreases in both productivity and product price. Suppose, for example, that mousetraps are sold in a competitive market, where their price is set by the interaction of supply and demand. Assume also that mousetrap production is subject to diminishing marginal returns. As more and more units of labor are added to a fixed quantity of plant and equipment, output expands by smaller and smaller increments.

Column 2 of Table 15.1 illustrates diminishing returns. The first laborer contributes a marginal product—or additional output—of six mousetraps per hour. From that point on, the marginal product of each additional laborer diminishes. It drops from five mousetraps to four to three and so on, until an extra laborer adds only one mousetrap to total hourly production.

The employer's problem, once production has reached the range of marginal diminishing returns, is to determine how many laborers to employ. She does so by considering the value of the marginal product of labor. Column 3 shows the market price of each mousetrap, which we will assume remains constant at $2. By multiplying that dollar price by the marginal product of each laborer (column 2) the employer arrives at the value of each laborer's marginal product (column 4). This is the highest amount that

she will pay each laborer. She is willing to pay less (and thereby gain profit), but she will not pay more.

TABLE 14.1 Computing the Value of the Marginal Product of Labor

Units of Labor (1)	Marginal Product of Each Laborer (per Hour) (2)	Price of Mousetraps in Product Market (3)	Value of Each Laborer to Employer (Value of the Marginal Product) [(2) x (3)] (4)
First laborer	6	$2	$12
Second laborer	5	2	10
Third laborer	4	2	8
Fourth laborer	3	2	6
Fifth laborer	2	2	4
Sixth laborer	1	2	2

If the wage rate is slightly below $12 an hour, the employer will hire only one worker. She cannot justify hiring the second worker if she has to pay him $12 for an hour's work and receives only $10 worth of product in return. If the wage rate is slightly lower than $10, the employer can justify hiring two laborers. If the wage rate is lower still—say, slightly below $4—the employer can hire as many as five workers.

Following this line of reasoning, we can conclude that the demand curve for mousetrap makers, like the demand curves for other goods, slopes downward. That is, the lower the wage rate, everything else held constant, the greater the quantity of labor demanded. Theoretically, what is true of one employer must be true of all. That is, the market demand curve for a given type of labor must also slope downward (see Figure 15.1).[1] Thus profit-maximizing employers will not employ workers if they have to pay them more, in wages and fringe benefits, than they are worth. What they are worth depends on their productivity and the market value of what they produce.

If the price of the product, mousetraps in this example, increases, the employer's demand for mousetrap makers will shift—say, from D_1 to D_2 in Figure 15.1. Because the market value of the laborers' marginal product has risen, producers now want to sell more mousetraps and will hire more workers to produce them. Look back again at Table 15.1. If the price of mousetraps rises from $2 to $4, the value of each worker's marginal product doubles. At a wage rate of $10 an hour, an employer can now hire as many as four workers. (Similarly, if the price of the final product falls below $2, the demand for workers will also fall.)

[1] The reader may get the impression that the market demand curve for labor is derived by horizontally summing the value of marginal product curves of individual firms, which are derived directly from tables like Table 15.1. Strictly speaking, that is not the case. However, these are refinements of theory that will be reserved for other, more advanced textbooks and courses.

When technological change improves worker productivity, the demand for workers may increase. If workers produce more, the value of their marginal product may rise, and employers may then be able to hire more of them. Such is not always the case, however. Sometimes an increase in worker productivity decreases the demand for labor. For instance, if worker productivity increases throughout the industry, rather than in just one or two firms, more mousetraps may be offered on the market, depressing the equilibrium price. The drop in price reduces the value of the workers' marginal product and may outweigh the favorable effect of the increase in productivity. In such cases the demand for labor will fall. Consumers will pay less, but employees in the mousetrap industry will have fewer employment opportunities and earn less .

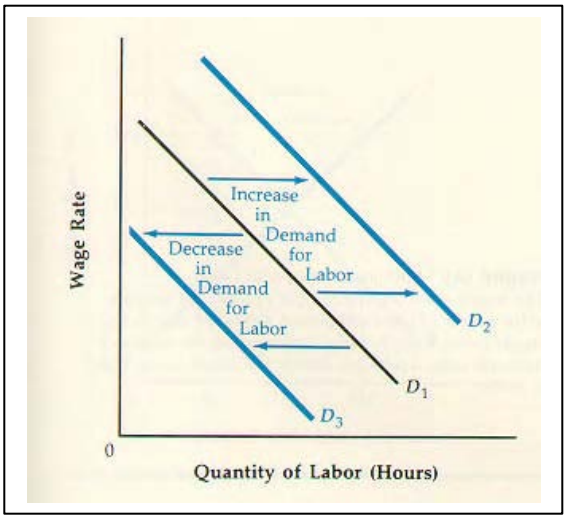


FIGURE 15.1 Shift in Demand for Labor

The demand for labor, like all other demand curves, slopes downward. An increase in the demand for labor will cause a rightward shift in the demand curve, from D_1 to D_2. A decrease will cause the leftward shift, to D_3.

The Supply of Labor

The **supply of labor** is the assumed positive relationship between the real wage rate and the number of workers (or work hours) offered for employment during a given period, everything else held constant. The supply curve for labor generally slopes upward. At higher wage rates, more workers will be willing to work longer hours than at lower wage rates (see Figure 15.2). If you survey your MBA classmates, for example, you will probably find that more of them would be willing to work at a job that paid $50 an hour than would work for $20 an hour. (At $500 an hour, most would be willing to work without hesitation, aside for a few lawyers and consultants!)

The supply of labor depends on the opportunity cost of a worker's time. Workers can do many different things with their time. They can use it to construct mousetraps, to do other jobs, to go fishing, and so on. Weighing the opportunity cost of each activity, the worker will allocate his time so that the marginal benefit of an hour spent doing one thing will equal the marginal benefit of time that could be used elsewhere. Because some kinds of work are unpleasant, workers will require a wage to make up for the time lost from leisure activities like fishing. To earn a given wage, a rational worker will give up the activities he values least. To allocate even more time to a job (and give up more valuable leisure-time activities), a worker will require a higher wage.

Given this cost-benefit tradeoff, employers who want to increase production have two options. They can hire additional workers or ask the same workers to work longer hours. Those who are currently working for $20 an hour must value time spent elsewhere at less than $20 an hour. To attract other workers, people who value their time spent elsewhere at more than $20 an hour, employers will have to raise the wage rate, perhaps to $22 an hour. To convince current workers to put in longer hours – to give up more attractive alternative activities – employers will also have to raise wage rates. In either case, the labor supply curve slopes upward. More labor is supplied at higher wages.

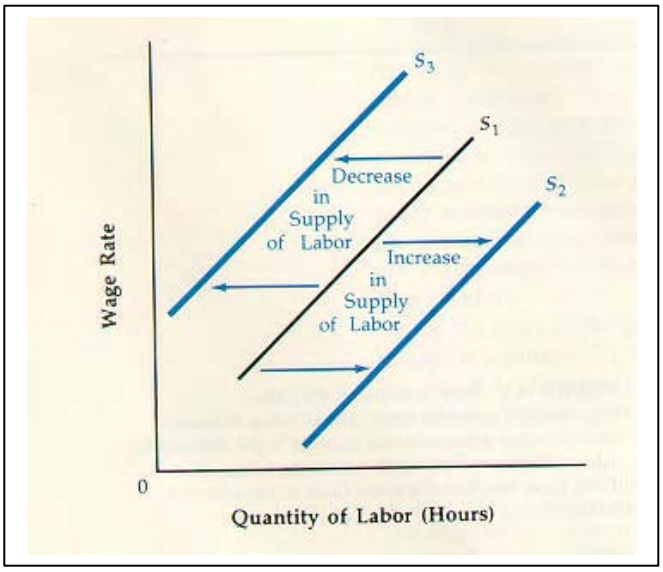


Figure 15.2 Shift in the Supply of Labor

The supply curve for labor slopes upward. An increase in the supply of labor will cause a rightward shift in the supply curve from S_1 to S_2. A decrease in the supply of labor will cause a leftward shift in the supply curve, from S_1 to S_3.

The supply curve for labor will shift if the value of employees' alternatives changes. For example, if the wage that mousetrap makers can earn in toy production goes up, the value of their time will increase. The supply of labor to the mousetrap industry should then decrease, shifting upward and to the left from S_1 to S_3, in Figure 15.2. This shift in the labor supply curve means that less labor will be offered at any given wage rate, in a particular labor market. To hire the same quantity of labor—to keep mousetrap makers from going over to the toy industry—the employer must increase the wage rate.

The same general effect will occur if workers' valuation of their leisure time changes. Because most people attach a high value to time spent with their families on holidays, employers who want to maintain operations then generally have to pay a premium for workers' time. The supply curve for labor on holidays lies above and to the left of the regular supply curve. Conversely, if for any reason the value of workers' alternatives decreases, the supply curve for labor will shift down to the right. If wages in the toy industry fall, for instance, more workers will want to move into the mousetrap business, increasing the labor supply in the mousetrap market.

Equilibrium in the Labor Market

A competitive market is one in which neither the individual employer nor the individual employee has the power to influence the wage rate. Such a market is shown in Figure 15.3. Given the supply curve *S* and the demand curve *D*, the wage rate will settle at W_1, and the quantity of labor employed will be Q_2. At that combination, defined by the intersection of the supply and demand curves, those who are willing to work for wage W_1 can find jobs.

The equilibrium wage rate is determined much as the prices of goods and services are established. At a wage rate of W_2, the quantity of labor employers will hire is Q_1, whereas the quantity of workers willing to work is Q_3. In other words, at that wage rate a surplus of labor exists. Note that all the workers in this surplus except the last one are willing to work for less than W_2. That is, up to Q_3, the supply curve lies below W_2. The opportunity cost of these workers' time is less than W_2. They can be expected to accept a lower wage, and over time they will begin to offer to work for less than W_2. Other unemployed and employed workers must then compete by accepting still lower wages. In this manner the wage rate will fall toward W_1. In the process, the quantity of labor that employers can afford to hire will expand from Q_1 toward Q_2.

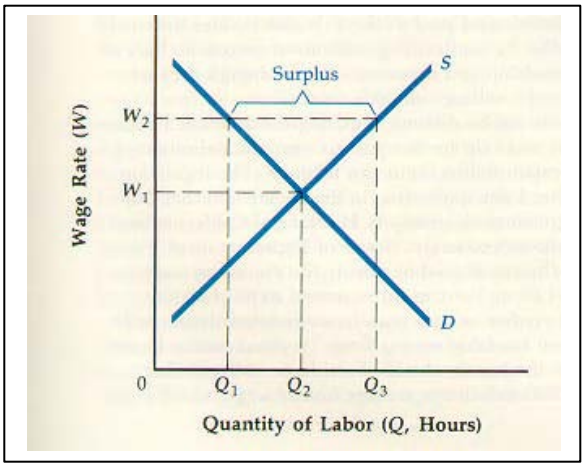


FIGURE 15.3 Equilibrium in the Labor Market

Given the supply and demand curves for labor *S* and *D*, the equilibrium wage will be W_1, and the equilibrium quantity of labor hired, Q_2. If the wage rate rises to W_2, a surplus of labor will develop, equal to the difference between Q_3 and Q_1.

Meanwhile, the falling wage rate will convince some workers to take another opportunity, such as going fishing or getting another job. As they withdraw from the market, the quantity of labor supplied will decline from Q_3 toward Q_2. The quantity supplied will meet the quantity demanded—and eliminate the surplus—at a wage rate of W_1.

In practice, the money wage rate—the number of dollars earned per hour—may not fall. Instead, the general price level may increase while the money wage rate remains constant. But the real wage rate—that is, what the money wage rate will buy—still falls, producing the same general effects: fewer laborers willing to work, and more workers demanded by employers. When economists talk about wage increases or decreases, they mean changes in the real wage rate, or in the purchasing power of a worker's paycheck.

Conversely, if the wage rate falls below W_1, the quantity of labor demanded by employers will exceed the quantity supplied, creating a shortage. Employers, eager to hire more workers at the new cheap wage, will compete for the scarce labor by offering slightly higher wages. The quantity of labor offered on the market will increase, but at the same time these slightly higher wages will cause some employers to cut back on their hiring. In short, in a competitive market, the wage rate will rise toward W_1, the equilibrium wage rate.

Why Wage Rates Differ

In a world of identical workers doing equivalent jobs under conditions of perfect competition, everyone would earn the same wage. In the real world, of course, workers differ, jobs differ, and various institutional factors reduce the competitiveness of labor markets. Some workers therefore earn higher wages than others. Indeed, the differences in wages can be inordinately large. (Compare the hourly earnings of Sylvester Stallone to those of elementary school teachers.) Wages differ for many reasons, including differences in the nonmonetary benefits (or costs) of different jobs. Conditions in different labor markets may differ in such a way as to cause wages to differ. Differences in the inherent abilities and acquired skills of workers can generate substantial differences in wages. Finally, discrimination against various groups often lowers the wages of people in those groups.

Differences in Nonmonetary Benefits

So far we have been speaking as if the wage rate were the key determination of employment. What about job satisfaction and the way employers treat their employees—are these issues not important? Some people accept lower wages in order to live in the Appalachians or the Rockies: college professors forgo more lucrative work to be able to teach, write, and set their own work schedules. The congeniality of their colleagues is another significant nonmonetary benefit that influences where and how much people work. Power, status, and public attention also figure in career decisions.

The tradeoffs between the monetary and nonmonetary rewards of work will affect the wage rates for specific jobs. The more importance people place on the nonmonetary benefits of a given job, the greater the labor supply. Added to wages, nonmonetary benefits could shift the labor supply curve from S_1 to S_2 in Figure 15.4, lowering the wage rate from W_2 to W_1. Even though the money wage rate is lower, however, workers are better off according to their own values. At a wage rate of W_1, their nonmonetary benefits equal the vertical distance between points *a* and *b*, making their full wage equal to W_3. The **full wage rate** is the sum of the money wage rate and the monetary equivalent of the nonmonetary benefits of a job.

Workers who complain they are paid less than workers in other occupations often fail to consider their full wages (money wage plus nonmonetary benefits). The worker with a lower monetary wage may be receiving more nonmonetary rewards, including

freedom from intense pressure, comfortable surroundings, and so on. The worker with the higher money wage may actually be earning a lower full wage than the worker with nonmonetary income. Certainly many executives must wonder whether their high salaries compensate them for their lost home life and leisure time, and teachers who envy the higher salaries of coaches should recognize that a somewhat higher wage rate is necessary to offset the increased risk of being fired that goes with coaching.

Employers can benefit from providing employees with nonwage benefits. A favorable working climate attracts more workers at lower wages. Although benefits can be costly, they are worthwhile as long as they lower wages more than they raise other labor costs. Some nonwage benefits, like air conditioning and low noise levels, also raise worker productivity. Needless to say, an employer cannot justify unlimited nonwage benefits. Employers will not pay more in wages, monetary or nonmonetary, than a worker is worth. In a competitive labor market they will tend to pay all employees a wage rate equal to the marginal value of the last employee hired.

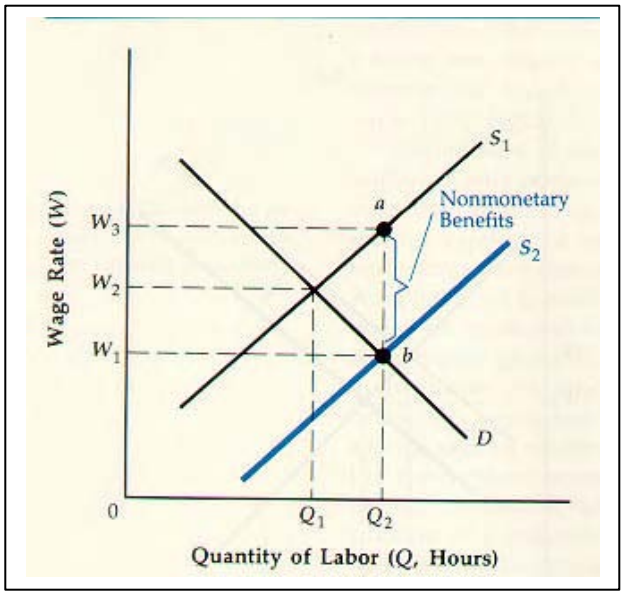


FIGURE 15.4 The Effect of Nonmonetary Rewards on Wage Rates

The supply of labor is greater for jobs offering nonmonetary benefits—S_2 rather than S_1. Given a constant demand for labor, the wage rate will be W_2 for workers who do not receive nonmonetary benefits and W_1 for workers who do. Even though wages are lower when nonmonetary benefits are offered, workers are still better off; they earn a total wage equal, according to their own values, to W_3.

Differences Among Markets

Differences in nonmonetary benefits explain only part of the observed differences in wage rates. Supply and demand conditions may differ between labor markets. As Figure 15.5 shows, given a constant supply of labor, *S*, a greater demand for labor will mean a higher wage rate. Conversely, given a constant demand for labor will mean a lower wage rate. Depending on the relative conditions in different markets, wages may—or may not—differ significantly.

People in different lines of work may also earn different wages because consumers value the products they produce differently. Automobile workers may earn more than textile workers because people are willing to pay more for automobiles than for clothing. Consumer preferences contribute to differences in the value of the marginal product of labor and ultimately in the demand for labor.

By themselves, relative product values cannot explain long-run differences in wages. Unless textile work offers compensating nonmonetary benefits, laborers in that industry will be attracted to higher wages elsewhere, perhaps in the automobile industry. The supply of labor in the automobile industry will rise, and the wage rate will fall. In the long run, the wage differential will decrease or even disappear.

Certain factors may perpetuate the money wage differential in spite of competitive market pressures. Textile workers who enjoy living in North or South Carolina may resist moving to Detroit, Michigan, where automobiles are manufactured. In that case, the nonmonetary benefits associated with textile work offset the difference in money wages. In addition, the cost of acquiring the skills needed for automobile work may act as a barrier to movement between industries—a problem we will address shortly.

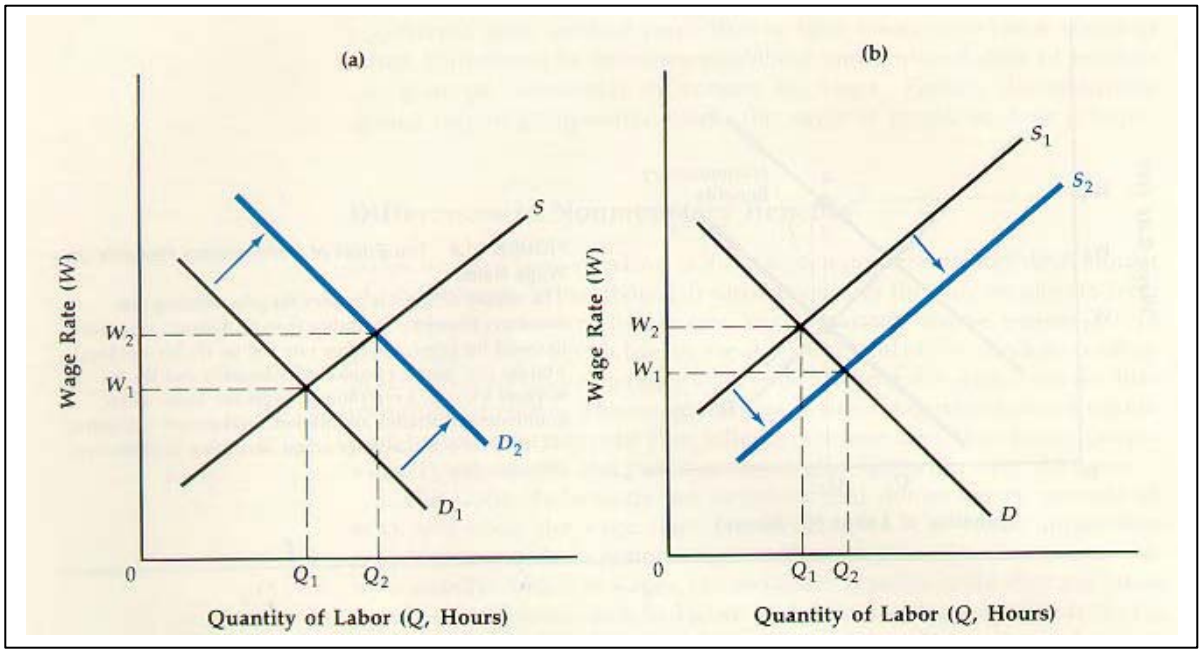


FIGURE 15.5 The Effect of Differences in Supply and Demand on Wage Rates

In competitive labor markets, higher demand for labor (D_2 in part (a)) will bring a higher wage rate. A higher supply of labor ($S2$ in part (b)) will bring a lower rate

Differences Among Workers

Differences in labor markets do not explain wage differences among people in the same line of work. Differences among workers must be responsible for that disparity. Some people are more attractive to employers. Employers must pay such workers more because their services are eagerly sought after, but they can afford to pay them more, because their marginal product is greater.

Mark McGuire earns an extremely high salary. The St. Louis Cardinals are willing to pay him so well both because of his popularity among fans -- when McGuire

plays, ballpark crowds are bigger—and because he is a successful hitter. Because a winning team generally attracts more support than a losing one. McGuire's presence indirectly boosts the team's earnings. In other words, McGuire is in a labor submarket like the one shown by curve D_2 in Figure 15.5(a). Other players are in submarket D_1.

Differences in skill may also account for differences in wages. Most wages are paid not just for a worker's effort but also for the use of what economists call human capital. **Human capital** is the acquired skills and productive capacity of workers. We usually think of capital as plant and equipment—for instance, a factory building and the machines it contains. A capital good is most fundamentally defined, however, as something produced or developed for use in the production of something else. In this sense capital goods include the education or skill a person acquires for use in the production process. The educated worker, whether a top-notch mechanic or a registered nurse, holds within herself capital assets that earn a specific rate of return. In pursuing professional skills, the worker, much like the business entrepreneur, takes the risk that the acquired assets will become outmoded before they are fully used. In the 1970's and 1980's students who majored in history expecting to teach found that their investment in human capital did not pay off. Many were unable to get jobs in their chosen field.

Finally, wage differences can result from social discrimination, whether sexual, racial, religious, ethnic, or political. Potential employees are often grouped according to some easily identifiable characteristic, such as sex or skin color. Employment decisions are then made primarily on the basis of the group to which the individual belongs, rather than on individual merit. Thus a qualified woman may not be considered for an executive job because women as a group are excluded. To the extent that employers prefer to work with certain groups, like whites or men, the labor market will be segmented. Employees in different submarkets, with different demand curves and wage differentials, will be unable to move easily from one market to another. The barriers to the free movement of workers allow wage differences to persist.

Competition among producers in the market for final goods can weaken (but not necessarily eliminate) discriminatory practices. Suppose employers harbor a deep-seated prejudice against women, which depresses the market demand and wage rates for female workers. If there is no rational reason for preferring men—if women are just as productive as men—an enterprising producer can hire women, pay them less, undersell the other suppliers, and take away part of their markets. Under competitive pressure, employers will start to hire women in order to keep their market shares. As a result, the demand for women workers will rise, while the demand for men will fall. Such competition may not eliminate the wage differential between men and women, but it can reduce it. In industries where employers exercise market power, social discrimination may persist.

Stricter Housing Standards for Migrants

In the 1960s, television news documentaries have publicized the substandard, even squalid housing commonly provided to migrant farm workers. Most housing for

migrants lacked plumbing and running water. Sleeping arrangements consisted of a few mattresses thrown on the floor.

To many, the obvious solution to the problem was to impose stricter housing standards on employers of migrant workers. Yet one consequence of such legislation had been reduced employment opportunities for migrant workers.[2] Figure 15.6 shows how the increased nonmonetary benefits lowered the demand for migrant labor. In a completely free market, employers are willing to pay a money wage of W_2 for migrant labor. If they are forced to pay workers more by meeting higher housing standards, their demand curve for migrant labor will fall from D_1 to D_2. The equilibrium money wage rate will fall to W_1, and employment opportunities will be reduced from Q_3 to Q_2. Again, as in the case of the minimum wage, those who keep their jobs may be better off. Their *full* wage rate will rise from W_2 to W_3 (W_1 in money wages plus $W_3 - W_1$ in nonmonetary benefits), but the workers who are not hired will suffer a loss in income.

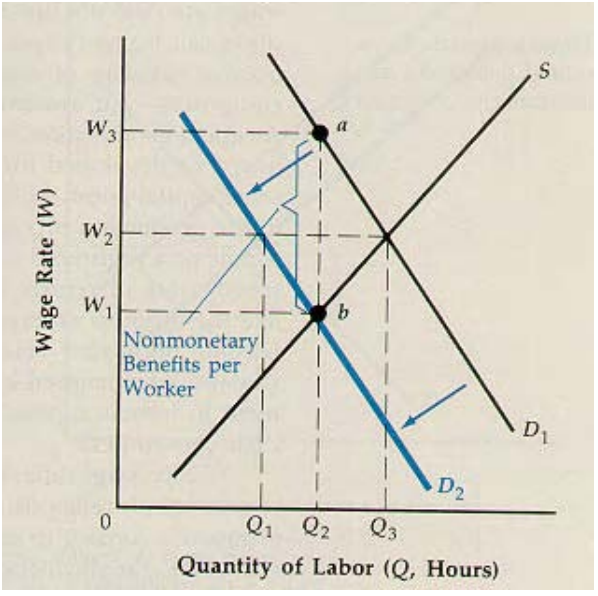


FIGURE 15.6 The Effect of Stricter Housing Standards on Employment

Higher housing standards for migrant workers will reduce employers' demand for migrant labor from D_1 to D_2. The money wage rate will fall from W_2 to W_1, but the nonmonetary benefits of improved housing will increase by the vertical distance between points *a* and *b*. Although workers will be earning a full wage of W_3, fewer of them—Q_2 instead of Q_3—will be hire

[2] Milton Friedman, "Migrant Workers," *Newsweek* (July 27, 1970): 60.

The farmers who employ migrant workers are caught in a competitive bind. Consumers want to buy their food at the lowest possible price. As producers, farmers must be able to sell their produce at a competitive price. That means minimizing the cost of production, including the full wage rate paid to employees. If farmers are forced to provide better housing for their laborers, they must reduce costs in other ways, including the substitution of machinery for labor. This is precisely what happened in many farm areas since the imposition of stricter housing standards. A federal law establishing migrant housing standards was passed in the late 1960s. In 1969 the farm labor service of the Michigan Employment Security Commission arranged jobs and housing for 27,163 migrants, but in the summer of 1970 it estimated that it would be able to place only 7,000 to 8,000 workers. State and local officials forecast that on balance, the new housing standards would eliminate 6,000 to 10,000 jobs. Meanwhile many growers, stung by the bad publicity surrounding migrant housing, closed their camps and switched to mechanical harvesting. As one grower put it, "It might be cheaper for me to continue using migrant help for a few more years, but mechanization is the trend of the future. And no matter what kind of housing I provide I'm going to be criticized for mistreating migrants. So I might as well switch now."[3]

Monopsonistic Labor Markets

Competition is bad for those who have to compete. Not only as producers but as employers, firms would rather control competitive forces than be controlled by them. They would like to pay employees less than the market wage—but competition does not give them that choice

Similarly, workers find that competition for jobs prevents them from earning more than the market wage. Thus doctors, truck drivers, and barbers have an interest in restricting competition in their labor markets. Acting as a group, they can acquire some control over their employment opportunities and wages.

Such power is difficult to maintain without the support of the law or the threat of violence, whether real or imagined. It comes at the expense of the consumer, who will have fewer goods and services to choose from at higher prices. As always, the exercise of power by one group leads not only to market inefficiencies but also to attempts by other groups to counteract it. The end result can be reduction in the general welfare of the community.

This section examines both employer and employee power in the labor market; the conditions that allow it to persist; its influence on the allocation of resources; and its effects on the real incomes of workers, consumers, and entrepreneurs.

---

[3] "Housing Dispute Spurs Migrant Farmers to Switch to Machines from Migrant Help," *Wall Street Journal*, June 29, 1970, p. 18

The Monopsonistic Employer

Power is never complete. It is always circumscribed by limitations of knowledge and the forces of law, custom, and the market. Within limits, employers can hire and fire, and can decide what products to produce and what type of labor to employ. Laws restrict the conditions of employment (working hours, working environment) they may offer, however, as well as their ability to discriminate among employees on the basis of sex, race, age, or religious affiliation. Competition imposes additional constraints. In a highly competitive labor market, an employer who offers very low wages will be outbid by others who want to hire workers. Competition for labor pushes wages up to a certain level, forcing some employers to withdraw from the market but permitting others to hire at the going wage rate.[4]

For the individual employer, then, the freedom of the competitive market is a highly constrained freedom. Not so, however, for those lucky employers who enjoy the power of a monopsony. A **pure monopsony** is the sole buyer of a good, service, or resource. (Monopsony should not be confused with monopoly, the single seller of a good and service.) The term is most frequently used to indicate the sole or dominant employer of labor in a given market. A good example of a monopsony would be a large coal-mining company in a small town with no other industry. A firm that is not a sole employer but that dominates the market for a certain type of labor is said to have monopsony power. **Monopsony power** is the ability of a producer to alter the price of a resource by changing the quantity employed. By reducing competition for workers' services, monopsony power allows employers to suppress the wage rate.

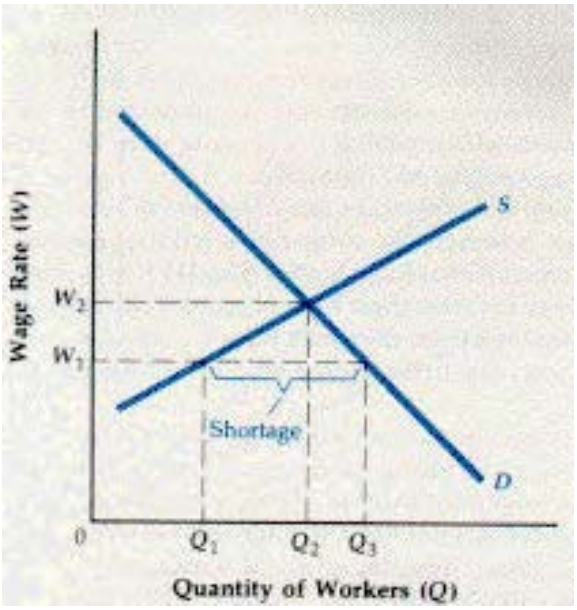


FIGURE 15.7 The Competitive Labor Market

In a competitive market, the equilibrium wage rate will be W_2. Lower wage rates, such as W_1, would create a shortage of labor, and employers would compete for the available laborers by offering a higher wage. In pushing up the wage rate to the equilibrium level, employers impose costs on one another. They must pay higher wages not only to new employees, but also to all current employees, in order to keep them.

[4] Competitors who do not hire influence the wage rate just as much as those who do; their presence on the sidelines keeps the price from falling. If firm lowers its wages, other employers may move into the market and hire away part of the work force.

The Cost of Labor

Monopsony power reduces the costs of competitive hiring. Assume that the downward-sloping demand curve *D* in Figure 15.7 shows the market demand for workers, and the upward-sloping supply curve *S* shows the number of workers willing to work at various wage rates. If all firms act independently—that is, if they compete with one another—the market wage rate will settle at W_2, and the number of workers hired will be Q_2. At lower wage rates, such as W_1, shortages will develop. As indicated by the market demand curve, employers will be willing to pay more than W_1. If a shortage exists, the market wage will be bid up to W_2.

An increase in the wage rate will encourage more workers to seek jobs. As long as there is a shortage, however, the competitive bidding imposes costs on employers. The firm that offers a wage higher than W_1 forces other firms to offer a comparable wage to retain their current employees. If those firms want to acquire additional workers, they may have to offer an even higher wage. As they bid the wage up, firms impose reciprocal costs on one another, as at an auction.

Because any increase in wages paid to one worker must be extended to all, the total cost to all employers of hiring even one worker at a higher wage can be staggering. If the wage rises from W_1 to W_2 in Figure 15.7, the total wage bill for the first Q_1 workers rises by the wage increase $W_2 - W_1$ times Q_1 workers. Table 15.2 shows how the effect of a wage increase is multiplied when it must be extended to other workers. The first two columns reflect the assumption that as the wage rate rises, more workers will accept jobs. If only one worker is demanded, he can be hired for $20,000. The firm's total wage bill will also be $20,000 (column 3). If two workers are demanded, and the second worker will not work for less than $22,000, the salary of the first worker must also be raised to $22,000. The cost of the second worker is therefore $24,000 (column 4): $22,000 for his services plus the $2,000 raise that must be given to the first worker.

TABLE 15.2 Market Demand for Tomatoes

Number of Workers Willing to Work (1)	Annual Wage of Each Worker (2)	Total Wage Bill [(1) times (2)] (3)	Marginal Cost of Additional Worker [Change in (3)] (4)
1	$20,000	$ 20,000	$20,000
2	22,000	44,000	24,000
3	24,000	72,000	28,000
4	26,000	104,000	32,000
5	28,000	140,000	36,000
6	30,000	180,000	40,000

The cost of additional workers can be similarly derived. When the sixth worker is added, she must be offered $30,000 and the other five workers must each be given a $2,000 raise. The cost of adding this new worker, called the marginal cost of labor, has risen to $40,000. The **marginal cost of labor** is the additional cost to the firm of expanding employment by one additional worker. Note that as the number of workers hired increases, the gap between the marginal cost of labor and the going wage rate expands. When two workers are hired, the gap is $2,000 ($24,000-$22,000). When six are employed, it is $10,000 ($40,000-$30,000).

Figure 15.8, based on columns 1 and 4 of Table 15.2, shows the marginal cost of labor graphically. The marginal cost curve lies above the supply curve, for the cost of each new worker hired (beyond the first worker) is greater than the worker's salary.

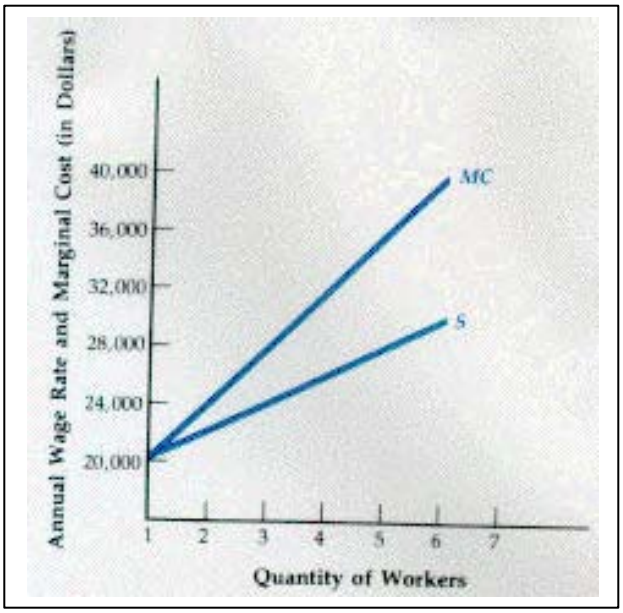


FIGURE 15.8 The Marginal Cost of Labor

The marginal cost of hiring additional workers is greater than the wages that must be paid to the new workers. Therefore the marginal cost of labor curve lies above the labor supply curve

The Monopsonistic Hiring Decision

The monopsonistic employer does not get caught in the competitive bind. By definition it is the only or dominant employer. Like a monopolist, the monopsonist can search through the various wage-quantity combinations on the labor supply curve for the one that maximizes profits. The monopsonist will keep hiring more workers as long as their contribution to revenues is greater than their additional cost, as shown by the marginal cost of labor curve *MC* in Figure 15.9. To maximize profits, in other words, the monopsonist will hire until the marginal cost of the last worker hired (*MC*) equals his marginal value, as shown by the textiles market demand curve for labor. Given the demand for labor *D*, the monopsonist's optimal employment level will be Q_2, where the marginal cost and demand for labor curves intersect. Note that that level is lower than the competitive employment level, Q_3.

Why hire where marginal cost equals marginal value? Suppose the monopsonist employed fewer workers—say Q_1. The marginal value of worker Q_1 would be quite high (point *a*), while her marginal cost would be low (point *b*). The monopsonist would be forgoing profits by hiring only Q_1 workers. Beyond Q_2 workers, the reverse would be true. The marginal cost of each new worker would be greater than his marginal value. Hiring more than Q_2 workers would reduce profits.

Once the monopsonist has chosen the employment level Q_2, it pays workers no more than is required by the labor supply curve, *S*. In Figure 15.9, the monopsonist must pay only W_1—much less than the wage that would be paid in a competitive labor market, W_2. In other words, the monopsonist hires fewer workers and pays them less than an employer in a competitive labor market.

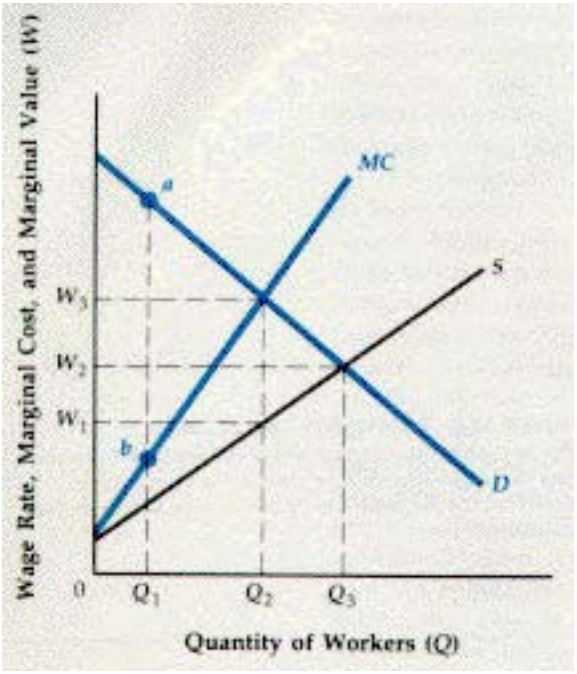


FIGURE 15.9 The Monopsonist

The monopsonist will hire up to the point where the marginal value of the last worker, shown by the demand curve for labor, equals his or her marginal cost. For this monopsonistic employer, the optimum number of workers is Q_2. The monopsonist must pay only W_1 for that number of workers—less than the competitive wage level, W_2.

It is the monopsonistic firm's power to reduce the number of workers hired that enables it to hold wages below the competitive level. In a competitive labor market, if one firm attempts to cut employment and reduce wages, it will not be able to keep its business going, for workers will depart to other employers willing to pay the going market wage. The individual firm is not large enough in relation to the entire labor market to exercise monopsony power. It therefore must reluctantly accept the market wage, W_2, as a given.

Employer Cartels: Monopsony Power through Collusion

Envying the power of the monopsonist, competitive employers may attempt to organize a cartel. A **employer cartel** is any organization of employers that seeks to restrict the number of workers hired in order to lower wages and increase profits.

The usual way of lowering employment is to establish restrictive employment rules that limit the movement of workers from one job to another. Such rules tend to reduce the demand for labor. In Figure 15.10, demand falls from D_1 to D_2. As a result, the wage rate drops, from W_2 to W_1, and employment falls, from Q_3 to Q_2. Although the method of limiting employment is different from that used in monopsony, the effect is the same. Whether the monopsonistic firm equates marginal cost with marginal value (shown by curve D_1) or the employer cartel reduces the demand for labor (to D_2), employment still drops to Q_2. In both cases workers earn a wage rate of W_1—less than the competitive wage.

One industry in which employers have tried to cartelize the labor market is professional sports. Owners of teams have developed complex rules governing the hiring of athletes. In the National Football League (NFL), for example, teams acquire rights to negotiate with promising college players through an annual draft. Once one team has drafted a player, no other team in the league can negotiate with him (unless he remains unsigned until the next year's draft). Teams can buy and sell draft rights as well as rights to players already drafted, but within leagues they are prohibited from competing directly with one another for players' services. Violations of these rules carry stiff penalties, including revocation of a team's franchise.

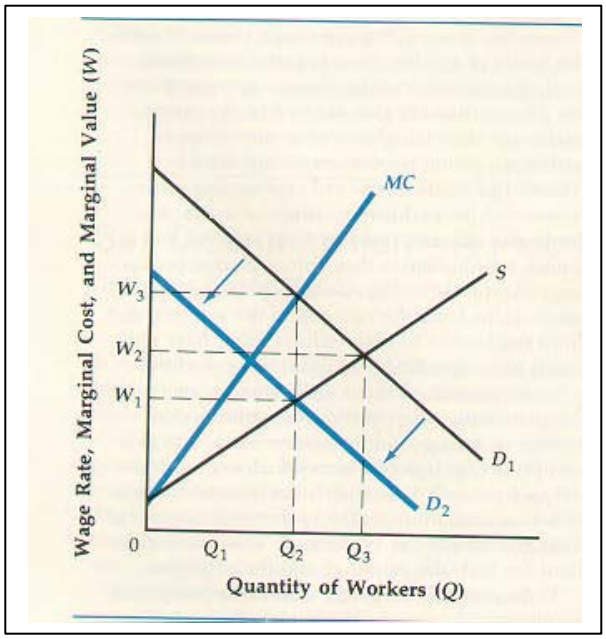


FIGURE 15.10 The Employer Cartel

To achieve the same results as a monopsonist, the employer cartel will devise restrictive employment rules that artificially reduce market demand to D_2. The reduced demand allows cartel members to hire only Q_2 workers at wage W_1—significantly less than the competitive wage, W_2.

MANAGER'S CORNER I: **Paying for Performance**

To this point in the chapter, our discussion has been focused on how labor "markets" work, and our interest has been on how the wage rate and other benefits are determined by the broad forces of supply and demand. However, markets must ultimately work with the interests of workers in mind. The problem most firms must solve is how to get

workers to do what they are supposed to do, which is work effectively and efficiently together for the creation of firm profits. This is no mean task, as we will see at various points in this book. There is a lot of trial and error in business, especially as it relates to how workers are paid. At the same time, thinking conceptually about the payment/incentive problem can help firms moderate the extent of errors in business.

One of the most fundamental rules of economics, and the *reason d'être* the discussions in the "Manager's Corners," is that if you offer people a greater reward, then they will do more of whatever is being rewarded, everything else equal. Many people find this proposition to be objectionable, because it implies that people can, to one degree or another, be "bought." Admittedly, incentives may not matter in all forms of behavior; some people will sacrifice their lives rather than forsake a strongly held principle. However, the proposition that incentives matter does seem to be applicable to a sufficiently wide range of behavior to be considered a "rule" that managers are well advised to keep in mind: Pay someone a higher wage -- such as time and a half -- and they will work overtime. Pay them double time, and they will even work holidays. There is some rate of pay at which a lot of people will work almost any time of the day or night on any day of the year.

This rule for incentives is not applicable only to the workplace. Parents know that one of the best ways to get their children to take out the garbage is to tie their allowance to that chore. Moreover, patients in psychiatric hospitals, many of whom have literally lost virtually all capacity for rational discourse, appear to respond to incentives. According to research, if mentally ill, institutionalized patients are paid for the simple tasks they are assigned (for example, sweeping a room or picking up trash), they will perform them with greater regularity.[5]

Even pigeons, well known for having the lowest form of birdbrains, respond to incentives. Granted, pigeons may never be able to grasp the concept of monetary rewards (offering them a dollar won't enlist much of a response), but pigeons apparently know how to respond to food rewards (offer a nut in the palm of your outstretched hand and a whole flock will descend, and maybe leave their mark, on your shoulder). From research, we also know that pigeons are willing to work -- measured by how many times they peck colored levers in their cages -- to get food pellets, and they will work harder if the reward for pecking is raised. Researchers have also been able to get pigeons to loaf on the job just like humans. How? Simply lower their rate of "pay."[6]

The "Right" Pay

It would appear that rules of incentives would lead managers everywhere to make sure that workers have the right incentives by always tying pay to some measure of performance. Clearly, the lone worker in a single proprietorship has the "right" incentive. His or her reward is the same as the reward for the whole firm. The full cost

[5] See Richard B. McKenzie and Gordon Tullock, The New World of Economics (New York: McGraw-Hill, 1994), chap. 4.
[6] Ibid.

of any shirking is borne by the worker/owner. However, such a congruence between the rewards of the owners and workers is nowhere else duplicated. There are always "gaps" between the goals of the owners and the workers, and the greater the number of workers, typically, the greater the gap in incentives. In very large firms, workers have greatly impaired incentives to pursue the goals of the owners. The workers are far removed from the owners by layers of bureaucracy, communications on firm goals are often imperfect, and each worker at the bottom of the firm pyramid can reason that his or her contributions to firm revenues and goals, or the lack of them, can easily go undetected. A reoccurring theme of this book is that when monitoring is difficult, one can expect many workers to exploit opportunities at their disposal.

And the opportunities taken can result in substantial losses in worker output. Management specialist Edward Lawler reported that during a strike at a manufacturing firm, a secretary was asked to take over a factory job and was paid on a piece rate basis. Despite no previous experience, within days she was turning out 375 percent more output than the normal worker who had spent 10 years on the job and was constantly complaining that the work standards were too demanding.[7] Obviously, the striking workers had been doing something other than working on the job.

How can managers improve incentives, reduce shirking, and increase worker productivity? At the turn of the century, the great management guru Frederick Taylor strongly recommended piece-rate pay as a means of partially solving what he termed the "labor problem," but he was largely ignored in his own time by both management and labor, and for the good reasons discussed in this chapter.[8]

There is a multitude of ways of getting workers to perform that don't involve money pay, and many of the ways are studied in disciplines like organizational behavior, which draws on the principles of psychology. Managers do need to think about patting workers on the back once in a while, clearly defining corporate goals, communicating goals in a clear and forceful manner, and exerting leadership.

Southwest Airlines, one of the more aggressive, cost-conscious, and profitable airlines, motivates its workers by creating what one analyst called a "community . . . resembling a 17th century New England town more than a 20th century corporation." The airline *bonds* its workers with such shared values as integrity, trust, and altruism.[9] But, a company with a productive corporate culture is almost surely a company with strong incentives in place to reward productivity. Without taking anything away from the corporate culture at Southwest Airlines, it should, however, be pointed out that one reason it has the lowest cost in the business is that its pilots and flight attendants are paid by the trip. This, along with a strong corporate culture, explains why Southwest's pilots and flight attendants hustle when the planes are on the ground. Indeed, Southwest has the shortest turn-around time in the industry. It pays the crews to do what they can to get

---

[7] Edward E. Lawler, III, Strategic Pay: Aligning Organizational Strategies and Pay Systems (San Francisco: Jossey-Bass Publishers, 1990), p. 58.
[8] Frederick W. Taylor, "A Piece Rate System," American Society of Mechanical Engineers Transactions, vol. 16 (1895), pp. 856-893.
[9] William G. Lee, "The New Corporate Republics," Wall Street Journal (September 26, 1994), p. 12.

their planes back in the air.[10] Motorola organizes its workers into teams and allows them to hire and fire their cohorts, determine training procedures, and set schedules. Federal Express' corporate culture includes giving workers the right to evaluate their bosses and to appeal their own evaluations all the way to the chairman. But, still, it's understandable why Federal Express delivery people move at least twice as fast as U.S. postal workers: FedEx workers have incentives to do so, whereas postal workers do not.[11]

We don't want to criticize the traditional, non-incentive methods for getting things done in business. Indeed, we have taken up the issue of "teams" discussed much earlier in the book, and the importance of virtues like "trust" will be raised before we conclude this chapter. At the same time, we wish to stress a fairly general and straightforward rule for organizing much production: *Give workers a direct detectable stake in firm revenues or profits in order to raise revenues and profits. Pay for performance*. One means of doing that is to make workers' pay conditional on their output: the greater the output from each worker, the greater the individual worker's pay.

Ideally, we should dispense with salaries, which are paid by the week or year, and always pay by the "piece" -- or "piece rate." Many firms -- for example, hosiery mills -- do pay piece rate; they pay by the number of socks completed (or even the number of toes closed). Piece rate can be expected to raise wages of covered workers for two reasons: First, the incentives can be expected to induce workers to work harder for more minutes of each hour and for more hours during the workday. Second, the piece-rate workers will be asked to assume some of the risk of production, which is influenced by factors beyond the workers' control. For example, how much each worker produces will be determined by what the employer does to provide workers with a productive work environment and what other workers are willing to do. So, piece-rate workers can be expected to demand and receive a *risk premium* in their paychecks. One study has, in fact, shown that a significant majority of workers covered under "output-related contracts" in the nonferrous foundries industry earn between 5 percent and 12 percent, depending on the occupation, more than their counterparts who are paid strictly by their time at work. Of that pay differential, about a fifth has been attributable to risk bearing by workers, which means that a substantial share of the pay advantage for incentive workers is attributable to the greater effort expended by the covered workers.[12]

However, such a rule – paying by the piece -- is hardly universally adopted. Indeed, piece-rate workers probably make up a minor portion of the total work force (we have not been able to precisely determine how prevalent piece-pay systems are). Many automobile salespeople, of course, are paid by the number of cars sold. Many lawyers are paid by the number of hours billed (and presumably services provided). Musicians are often paid by the number of concerts played.

---

[10] Howard Banks, "A Sixties Industry in a Nineties Economy," Forbes, May 9, 1994: pp. 107-112.

[11] FedEx actually tracks its delivery people on their routes, and the workers understand that their pay is tied to how cost-effective they are in their deliveries. Postal workers understand that they are not being so carefully monitored, mainly because there are no stockholders who can claim the profits from a speed-up in their work.

[12] Tron Petersen, "Reward Systems and the Distribution of Wages," Journal of Law, Economics, and Organizations, vol. 7 (special issue), 1991, pp. 130-158.

But there are relatively few workers in manufacturing and service industries whose pay is directly tied to each item or service produced. Professors are not paid by the number of students they teach. Office workers are not paid by the number of forms processed or memos sent. Fast food workers are not paid by the number of burgers flipped. Most people's pay is, for the most part, directly and explicitly tied to time on the job. They are generally paid by the hour or month or even year.

Admittedly, the pay of most workers has some indirect and implicit connection to production. Many workers know that if they don't eventually add more to the revenues of their companies than they take in pay, their jobs will be in considerable jeopardy. The question we find interesting is why "piece rate" -- or "pay for performance" -- is not a more widely employed pay system, given the positive incentives it potentially provides.

Many explanations for the absence of a piece-rate pay system are obvious and widely recognized.[13] The output of many workers cannot be reduced to "pieces." In such cases, no one should expect pay to be tied to that which cannot be measured with tolerable objectivity. Our work as university professors is hard to define and measure. In fact, observers might find it hard to determine when we are working, given that while at work, we may be doing nothing more than staring at a computer screen or talking with students in the hallways. Measuring the "pieces" of what secretaries and executives do is equally, if not more, difficult.

If a measure of "output" is defined when the assigned tasks are complex, the measure will not likely be all-inclusive. Some dimensions of the assigned tasks will not be measured, which means that workers' incentives may be grossly distorted. They may work only to do those things that are defined and measured -- and related to pay -- at the expense of other parts of their assignments. If workers are paid by the number of parts produced, with the quality of individual parts not considered, some workers could be expected to sacrifice quality in order to increase their production count. If professors were paid by the number of students in their classes, you can bet they would spend less time at research and in committee meetings (which would not be all bad). If middle managers were paid solely by units produced, they would produce a lot of units with little attention to costs. There is an old story from the days before the fall of communism in the former Soviet Union. According to the story, the managers of a shoe factory were given production quotas for the number of shoes they had to make, and they were paid according to how much they exceeded their quota. What did they do? They produced lots of shoes, *but only left ones!*

Much work is the product of "teams," or groups of workers, extending, at times, to the entire plant or office. Pay is often not related to output because it may be difficult to determine which individuals are responsible for the "pieces" that are produced.

---

[13] For a review of arguments offered by psychologists against incentive pay plans, see Alfie Kohn, "Why Incentive Plans Cannot Work," Harvard Business Review September-October 1993, pp. 54-63. Kohn sums up his argument, "Do rewards motivate people? Absolutely. They motivate people to get rewards" (p. 62), suggesting that the goals of the firm might not be achieved in the process, given the complexity of the production process and the margins workers can exploit. Kohn's criticisms are reviewed and critiqued in the last chapter of this book.

Because we took up the problems of forming and paying teams in an earlier chapter, we only remind readers that team production creates special incentive problems. Making the teams "small" is one way to enhance incentives by making the contributions, or lack thereof, of each team member noticeable to others on the team.

When workers are paid by salary, they are given some assurance that their incomes will not vary with firm output, which can go up and down for many reasons, not under their control. For example, how many socks a worker can stitch at the toe is dependent upon the flow of socks through the plant, over which the workers who do the stitching may have no control. When workers are paid by the piece, they are, in effect, asked to assume a greater risk that shows up in the variability of the income they take home. Granted, piece rate may give the workers a higher *average* income. However, in order for the piece rate system to work -- and be profitable for the firm -- the increase in expected worker productivity would have to exceed the *risk premium* that risk averse workers would demand. *Piece rate (or any other form of incentive compensation) is not employed in many firms simply because the risk premium workers demand is greater than their expected increase in productivity*. This is often the case because workers tend to be risk averse (or reluctant to take chances, or assume the costs associated with an uncertain and variable income stream).

If paid by the work done, workers would also have to worry about how changes in the general economy would affect their workloads and production levels. A downturn in the economy, due to forces that are global in scope, can undermine worker pay when pay is tied to output. When Du Pont introduced its incentive compensation scheme for its fibers division in 1988 -- under which a portion of the workers' incomes could be lost if profit goals were not achieved and could be multiplied if profit goals were exceeded – the managers and employees expected, or were told to expect, substantial income gains. [14] However, when the economy turned sour in 1990, employee morale suffered as profits fell and workers were threatened with reduced incomes. The incentive program was cancelled before the announced three-year trial period was up.[15] Du Pont obviously concluded that it could buy back worker morale and production by not subjecting worker pay to factors that were beyond worker control. Each individual employee could reason that there was absolutely nothing he could do about the national economy or, for that matter, about the work effort expended by the 20,000 other Du Pont workers who were covered by the incentive program. They could rightfully fear that their incomes were being put at risk by the free riding of all other workers.

This line of analysis leads to the conclusion that piece-rate (and other forms of incentive) pay schemes will tend to be used in firms where the risk to workers is relatively low (relative to the benefits of the improved incentives). This means that they will tend to be used where production is not highly variable and where, in the absence of piece-rate pay, workers can easily exploit opportunities to shirk. That is, they will tend to be used where workers cannot be easily monitored. For example, salespeople who are

[14] L. Hayes, "All Eyes on Du Pont's Incentive Program," Wall Street Journal, December 5, 1988, p. B1.
[15] R. Koening, "Du Pont Plan Linking Pay to Fibers Profit Unravels," Wall Street Journal, October 25, 1990, p. B1.

always on the road (which necessarily means that no one at the home office knows much about what they do on a daily basis) will tend to be paid, at least in part, by the "piece," in some form or another, say, by the sale.

Piece-rate pay systems may also be avoided because employers are likely to be in a better position to assume the risk of production variability than their employees are. This is because much of the variability in the output of *individual* workers will be "smoothed out" within a whole *group* of employees. When one worker's output is down, then another worker's output will be up. Workers will, in effect, be able to buy themselves out of the risk. If each of the workers sees the risk cost of the piece-rate system at $500 and the employer sees the risk cost at $100, then each worker can agree to give up, say, $110 in pay for the rights to a constant income. The worker gains, on balance, $390 in non-money income ($500 in risk cost reduction minus the $110 reduction in money wages). The employer gives up the piece-rate system simply because he or she can make a profit -- $10 in this example -- off each worker ($110 reduction in worker money wages minus the $100 increase in risk cost). *One would therefore expect, other things equal, piece-rate pay schemes would be more prevalent in "small" firms than in "large" ones.* Large employers are more likely to be able to smooth out the variability.

Also, piece-rate pay systems can only be used when and where employers can make credible commitments to their workers to abide by the pay system that they establish and not to cut the *rate* in the *piece-rate* when the desired results are achieved. Unfortunately, all too often managers are unable to make the credible commitment for the same reason that they might find, in theory, the piece-rate system to be an attractive way (in terms of worker productivity and firm profits) to pay workers. The basic problem is that both workers *and* managers have incentives to engage in opportunistic behavior to the detriment of the other group.

Managers understand that many workers have a natural inclination to shirk their responsibilities, to loaf on the job and misuse and abuse company resources with the intent of padding their own pockets. Managers also know that if they tie their workers' pay to output, then output may be expected to expand. Fewer workers will exploit their positions and loaf on the job. At the same time, the workers can reason that incentives also matter to managers. Like workers, managers are not always angels (and are sometimes outright devils, just like their workers) and can be expected, to one degree or another, to exploit their positions, achieving greater personal and firm gains at the expense of their workers.

Hence, workers can reason that if they respond to the incentives built into the piece-rate system and produce more for more pay, then managers can change the deal. The managers can simply raise the number of pieces that the workers must produce in order to get the previously established pay, or managers can simply dump what will then be excess workers.

To clarify this point, suppose a worker is initially paid $500 a week, and during the course of the typical week, he or she produces 100 pieces -- for an average pay of $5 per piece. Management figures that the worker is spending some time goofing off on the

job and that the worker's output can be raised if he or she is paid $5 for each piece produced.

If the worker responds by increasing his output to 150 pieces, the management can simply lower the rate to $3.50 per piece, which would give the worker $525 a week and would mean that the firm would take the overwhelming share of the gains from the worker's -- not management's -- greater efforts. The worker would, in effect, be working harder and more diligently with little to show for what he or she has done. By heeding the piece-rate incentive, the worker could be inadvertently establishing a higher production standard.

These threats are real. Managers at a General Motors panel stamping plant in Flint, Michigan announced that the company would allow workers to leave after they had satisfied daily production targets. Workers were soon leaving by noon. Management responded by increasing production targets. The result was a bitter workforce.[16]

So, one reason piece-rate systems aren't more widely used is that the systems can be abused by managers, which means that workers will not buy into them at reasonable rates of pay.

Another way of explaining the lack of use of piece-rate pay is that they often don't work as might be expected. Incentives still matter. The problem is that the much talked-about incentives are not there, or workers don't believe they are there. And workers don't believe the incentives are present because they don't -- or can't -- believe that their managers will resist the temptation to gain at their -- the workers' -- expense. Managers are unable to make what we have, in other contexts, called *a credible commitment* (or a position on which workers can rely), meaning they have not been able to convince their workers that they will not take advantage of them (just as the workers may have been taking advantage of their managers).

Indeed, the piece-rate system can have the exact opposite effect of the one intended. We have noted that workers can reason that their managers will increase the output demands if they produce more for any given rate. However, the implied relationship between output and production demands should also be expected to run the other way: That *is, the workers can reason that if managers will raise the production requirements when they produce more in response to any established rate, then managers should be willing to lower the production requirements when the workers lower their production after the piece-rate system is established.* Hence, the establishment of the piece-rate system can lead to a reduction in output as workers cut back on production. The purpose of the incentive pay may be to increase production, but the result can be to induce lower production standards for the same rate of pay. The workers' expectation can be that the rate of pay will be raised.

How? Suppose that the worker responds to the rate of $5 per piece by actually cutting back his or her total production from 100 to 75 pieces per week. Then management might be expected to increase the rate to, say, $6.50 per piece, leaving the

[16] See Benjamin Klein, Robert Crawford, and Armen Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," Journal of Law and Economics, vol. 21 (1978), pp. 297-326.

worker with $487.50 for the week, or a 2.5 percent reduction in pay for a 25 percent reduction in effort.

The lesson of this discussion is not that piece-rate pay incentives can't work. Rather, the lesson is that getting them right can be tricky. Managers must convincingly *commit* themselves to holding to the established piece rate and not exploiting the workers. The best way for managers to be believable is to create a history of living up to their commitments, which means creating a valuable reputation with their workers.

Lincoln Electric, a major producer of arc-welding equipment in Cleveland, makes heavy use of piece-rate pay. The system has resulted in a doubling of worker productivity since 1945 and continues to be successful for several reasons:

- First, the company has a target rate of return for shareholders, with deviations from that target either adding to or subtracting from their workers' year-end bonuses, with the bonus often amounting to 100 percent of workers' base pay.
- Second, employees largely own the firm, a fact that reduces the likelihood that piece rates will be changed.
- Third, management understands the need for credible commitments. According to one manager, "When we set a piecework price, that price cannot be changed just because, in management's opinion, the worker is making too much money . . .. Piecework prices can only be changed when management has made a change in the method of doing that particular job and under no other conditions. If this is not carried out 100 percent, piecework cannot work."[17]
- Fourth, Lincoln pursues a permanent employment policy. Permanent employees are guaranteed only 75 percent of normal hours, and management can move workers into different jobs in response to demand changes. Also, workers have agreed to mandatory overtime when demand is high (meaning that the firm doesn't have to hire workers in peak demand periods). In other words, workers and management have agreed to share some of the risk.
- Fifth, to combat quality problems, each unit produced is stenciled with the initials of the workers who produced it. If a unit fails after delivery because of flaws in production, the responsible workers can lose as much as 10 percent of their annual bonus.
- Sixth, large inventories are maintained to smooth out differences in the production rates of different workers.

Does this mean that managers can never raise the production standard for any given pay rate? Of course not. Workers should only be concerned if the standard is changed because of something *they* -- the workers -- did. If management in some way increases the productivity of workers (for example, introduces computerized equipment or rearranges the flow of the materials through the plant), independent of how much

[17] As quoted in Gary J. Miller, Managerial Dilemmas: The Political Economy of Hierarchy (New York: Cambridge University Press, 1992), p. 117.

effort workers apply, then the standard can be raised. Workers should not object. They are still getting their value for their effort. They are not being made worse off. What managers must avoid doing is changing the foundations of the work and then taking more in terms of a lower *pay rate* than they are due, which effectively means violating the contract or commitment with their workers.

As the Lincoln Electric manager notes: "Piecework prices can only be changed when management has made a change in the method of doing that particular job and under no other conditions."[18] Otherwise, piece-rate pay can have the exact opposite effect of the one intended.

Two-Part Pay

There are innumerable ways of paying people to encourage performance. The two-part pay contract -- salary plus commission -- is obviously a compromise between straight salary and straight commission pay structures. For example, a worker for a job placement service can be paid a salary of $1,500 a month, plus 10 percent of the fees received for any placement. If the recruiter can be expected to place one worker a month and the placement fee is $10,000, the worker's expected monthly income is $2,500 ($1,500 plus 10 percent of $10,000).

This form of payment can be mutually attractive to the placement firm and its recruiters because it accomplishes a couple of important objectives. First, the system can be a way by which workers and their employers can share the risks to reflect the way the actual placements depend on the actions of both the workers and their employers. While each worker understands that his or her placements are greatly affected by how hard and smart he or she personally works, each also knows that often, to a nontrivial degree, the placements are related to what all other workers and the employer do. Worker income is dependent on, for example, how much the employer advertises, seeks to maintain a good image for the firm, and develops the right incentives for *all* workers to apply themselves.

Workers have an interest in everyone in the firm working as a team, just as the employer does. Productive work by all can increase firm output, worker pay, and job security. As a consequence, while each worker may, in one sense, "prefer" all income in the form of a guaranteed fixed monthly check, the worker also has an interest in commission pay -- *if everyone else is paid commission and if perverse incentives are avoided*. Often each worker's income is dependent upon how hard others work. Individual recruiters, to carry forward our example, often benefit from the attempts of other recruiters to make successful, quality placements. Such efforts can spread and enhance the name of the firm, making it easier for all other recruiters to make placements.

Hence, a pay system that is based, to a degree, on commission can raise the incomes of all recruiters. Put another way, to the extent that one worker's income is

[18] Ibid.

dependent upon other workers' efforts, we should expect workers to favor a pay system that incorporates strong production incentives for all workers.

Of course, workers want an employer who can be trusted. They don't want to be caught in a situation in which the incentive system undercuts production, as we have suggested could be the case. As a consequence, workers favor bosses who are paid a premium because they can be trusted. They certainly do not want bosses who engage in opportunism by cutting the *rate* of pay when workers respond to incentive pay by working harder and increasing output.

Some combination of straight salary and piece-rate pay can achieve *optimum* incentives and, therefore, can maximize firm output, worker pay, and job security. We should not expect that maximum incentives are always achieved with pay tied strictly to production. Unfortunately, we can't say exactly what the combination should be. There is no one ideal pay combination, mainly because conditions of production -- including the actual contributions by different workers and the degree of trust -- vary so greatly across firms and industries. Our central point is that the two-part – or salary-plus-commission -- pay systems can help workers by aligning their interests to those of fellow workers and their employers, and do so without exposing workers to excessive risk.

With the two-part pay system, workers are given some security in that they can count on, for some undetermined amount of time, a minimum income level -- $1,500 in our example. The workers shift some of their risk to their employer, but the risk assumed by the employer need not equal the sum of the risk that the workers avoid. This is because, as noted earlier, the employer usually hires a number of people, and the variability of the income of the employer is, therefore, not likely to be as great as the variability of the individual workers' income. As noted, each worker should be willing to give up some higher average expected income, for example, $2,000 a month, when his or her income is totally dependent upon placements (under which system the commission might have to be 25 percent of the placement fee). Both parties gain, and both parties can see the pay system as a means of "incentivizing" the other.

The workers, in other words, may want to give up something in straight commission income in order that their employer will assume some of the risk but, possibly just as importantly, the employer will have an interest in facilitating (to the extent possible) the placement process. After all, with the monthly salary hanging over the employer's head, the employer will want to work to make sure that the workers can earn their monthly keep. Each month *some* workers might do poorly, but other workers can have offsetting experiences. Moreover, with the employer assuming some of the risk, the employer can be expected to work harder in the interest of the workers, reducing some of the remaining risk that the workers must assume. The net effect of the two-part pay system should be that both parties could gain precisely because each party is motivated to contribute to the success of the other.

Workers will also understand that if everyone has an incentive to work harder, then there will be greater production from their "team" effort, resulting in greater production, more profits, and greater job security (as well as more pay and fringe

benefits). Workers can also reason that some incentive pay can reduce the risk cost that the firm must incur, thus, once again, potentially improving everyone's well being.

Also, workers can surely understand the press of market competition. If their firm doesn't find ways of sharing and reducing risk and increasing worker output, then other firms in their markets surely will. That fact can spell market failure for firms and their workers who fail to adopt two-part pay systems, if they are mutually beneficial.

Incentive Pay Equals Higher Pay

Of course, firms can expect that incentive schemes that enhance firm profits do not come free of charge. According to one early study, the nearly 200 punch-press operators in Chicago who were paid piece rate earned, on average, 7 percent more than workers who did much the same jobs but who were paid a straight salary (so much per unit of time, for example, hour, week, or month).[19] According to another study involving more than 100,000 workers in 500 manufacturing firms within two industries, the incomes of the footwear workers on some form of piece-rate or salary-plus-commission pay averaged slightly over 14 percent more than the workers on salaries (with the differential ranging up to 31 percent for certain types of jobs). The workers in the men's coats and suits industry on piece rate averaged between 15 and 16 percent more than the salaried workers.[20] And the best evidence available suggests that the more workers' incomes are based on incentive pay, the greater the income differential between those who are earn piece-rate pay (or any other form of incentive pay) and those who don't.

Of course, it may be that the income differential between incentive-paid and salaried workers is a matter of the difference in the demands of the jobs incentive-paid workers and salaried workers take. Incentive-paid jobs may pay more because they are the jobs the most competent workers are most anxious to take. However, the studies cited have attempted to either look at incentive-paid and salaried workers in comparable jobs or have adjusted (by statistical, econometric means) the pay gaps for differences in the "quality" of the different jobs.[21]

One of the more obvious explanations for why incentive-paid workers earn more than salaried workers is that the incentive-paid workers accept more risk. After all, the incomes of the incentive-paid workers can vary not only with the workers' effort, but also with the promotional efforts of their firms and general economic conditions in the market, among a host of other factors. A firm's ad campaigns can complement a worker's efforts to sell a product or service. A downturn in the national economy can make selling more

[19] J. H. Pencavel, "Work Effort, On the Job Screening, and Alternative Methods of Remuneration," Research in Labor Economics (Greenwich, Conn.: JAI Press, 1977), pp. 225-259.
[20] Eric Seiler, "Piece-Rate Vs. Time-Rate: The Effect of Incentives on Earnings," Review of Economics and Statistics, vol. 66, no. 3 (1984), pp. 363-375.
[21] The study by Pencavel ("Work Effort, On the Job Screening, and Alternative Methods of Remuneration") adjusts the worker data for differences in education, experience, race, and union status. The second study by Seiler ("Piece -Rate Vs. Time-Rate: The Effect of Incentives on Earnings") adjusts for differences in union status, gender, location of employment, occupation, type of product, and method of production, among other variables.

difficult, effectively dropping the workers' rates of pay per hour (albeit for a long or short period of time). The incentive-paid workers' greater average pay amounts to a risk premium intended to account for the prospects that income may not always match expectations.

The business lesson is simple: To get workers to accept incentive pay, employers have to raise the pay. If both incentive-paid and salaried jobs were paid the same, workers would crowd into the salaried jobs, increasing the number of workers available to work for salaries and reducing the number of workers available to workers on commission. The incomes of the salaried workers, everything else being equal, would tend to fall, while the incomes of the incentive-paid workers would tend to rise. If there were no considerations other than risk under the different pay schemes, the wage differential would continue to widen until the income difference were about equal to the difference in the added "risk cost" the incentive-paid workers suffered. That is to say, if the risk cost (or premium) were deducted from the pay of incentive-paid workers, the resulting net pay of the incentive-paid workers would be about the same as the pay of salaried workers.

But risk doesn't explain the entire differential (and would not ever likely do so). One of the studies mentioned at the start found that the "risk premium" accounted for only a little more than 3 percent of the pay differential in the footwear industry and only 6 percent of the difference in men's clothing (with a great deal of variance reported across occupational categories).[22] Another important portion of the differential can be explained by the dictum that is central to all Manager's Corners: Incentives matter! Incentive-paid workers simply gain more from extra work than do their salaried counterparts. A salaried worker is no doubt required to apply a given, minimal level of effort on the job. Salaried workers can choose to work more and produce more for the company. Their extra work might have some reward, a future raise or promotion, but such prospects are never certain. Many workers believe, with justification, that their raises are more directly tied to the number of years they survive at their firms than on how much extra they work and produce.

By way of contrast, the rewards of incentive-paid workers are much more immediate, direct, and contractual. Incentive-paid workers know that if they produce or sell more for their firms, their incomes will rise immediately and by a known amount. Accordingly, they have a greater incentive to apply themselves. One study in the early 1960s found that incentive pay improved worker productivity by as much as 40 percent, not all of which, as will be argued, is necessarily due to extra effort.[23]

Incentive pay does more than just motivate greater effort. Different methods of pay are likely to attract different workers. Workers who are relatively unproductive, or who just don't want to compete aggressively, are likely to opt out of incentive-paid work. They will tend to crowd in salaried jobs, where many other relatively unproductive and less aggressive workers are. In short, workers who tend to be more productive than

[22] Seiler, "Piece -Rate Vs. Time-Rate: The Effect of Incentives on Earnings."

[23] See G. L. Mangum. "Are Wage Incentives Becoming Obsolete?" Industrial Relations, vol. 2 (October 1962), pp. 73-96.

average can be expected to self-select into jobs with incentive pay. We should expect some firms to use incentive pay elements in many jobs simply to cull out the unproductive workers. Incentive pay allows job applicants who know that they are willing to work hard to convincingly communicate this willingness to prospective employers by their willingness to accept the challenge of incentive pay.

Of course, it should follow that the demands of the incentive-paid work -- and the resulting curb in the supply of incentive-paid workers -- will press the output and wages of incentive-paid workers up. At the same time, the crowding of less aggressive workers in salaried jobs will tend to increase the supply of salaried workers and lower their wages (if not absolutely, then certainly relative to incentive-paid workers).

If business becomes more uncertain, less predictable -- as many seem to think it has over the last couple of decades with the growing complexity and globalization of business -- we would expect the income gap between incentive-paid and salaried workers to widen. Employers will want to increase their competitive positions by giving their workers a greater incentive to work harder and smarter. Employers will want to shift a share of the growing business risk to their workers, at a price, of course, through greater reliance on commissions. At the same time, relatively speaking, more workers might seek to avoid the greater risk by trying to move to salaried jobs. However, their efforts will simply hold salaries down, widening the gap between incentive-paid and salaried jobs.

Those who have been willing to accept and cope with risks have seen their incomes rise. Those who have sought to stay on salaries have probably had to concede to accepting relatively (if not absolutely) lower wages. Growing business risk is surely not the only source of the expanding pay gap, but it is certainly one that has played a role.

To this point in the chapter, one of our more important conclusions has been that one of the reasons employers should pay workers in two parts -- in part by salary and in part by some form of tie to performance – is that both employer and employee can gain. The employer can accept this risk associated with having to meet a regular, contracted salary payment, and the employee can want the salary because it reduces his or her risk and, at the same time, gives the employer incentive to work hard at keeping the work going (in order that the salary can be met with relative ease). By adding to the fixed salary, the employer may curb the incentive the employer has to work hard and smart, but still the salary component can be a paying proposition for the employer because the overall compensation demands of the employee can fall by more than performance does. Similarly, the employee can lose more in "risk cost" than it loses in total compensation. Everyone can be happy, which is the sort of outcome managers should always seek.

However, for all its elegance, our discussion sidesteps a problem that managers must face when they are thinking about paying for performance: getting the workers to deal honestly when their pay is at stake. For example, consider the manager who has to deal with a sales force that works out in the "field," far removed from headquarters. The sales people are hard to monitor. They know a great deal more about their territories in terms of sales potential than the managers back at headquarters. How do the managers get the sales people to reveal the sales potential of their districts? This question is

especially troublesome when the sales people know that their revealed information will affect their sales performance criteria and the combination of the salary and commission components of their compensation package. If the manager at headquarters simply asks the sales people how much they can sale in their areas, there's a good chance the sales people will understate the sales potential. After all, some understatement harbors the potential of raising the salary and commission rate.

There is a simple solution that will encourage the sales people to deal honestly. The manager should offer the sales people a menu of combinations of salary and commission rates. Consider the set of three salary-commission rate combinations illustrated in Figure 15.11, which has pay on the vertical axis and sales on the horizontal axis. One pay package has a high salary, S_1 and a low commission rate, which is described by the low slope of the straight upward sloping compensation line that emerges from S_1 on the vertical axis. Another pay package has a salary component of S_2 and a higher commission rate, and yet a third has an even lower salary, S_3, and an even higher commission rate.

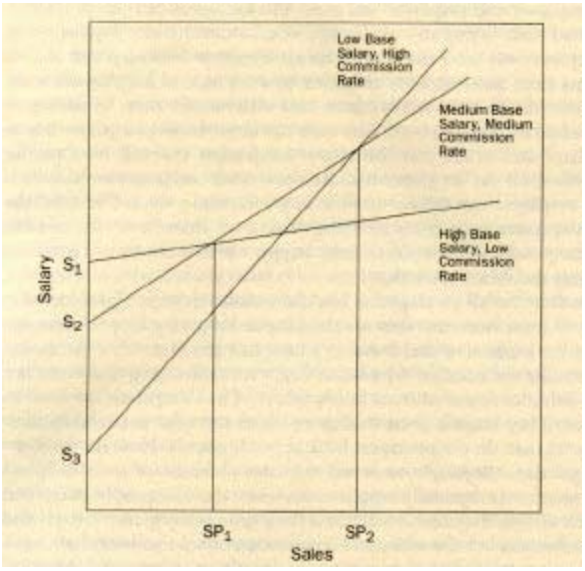


Figure 15.11 Menu of Two-Part Pay Packages

By varying the base salary and the commission rate, employers can get sales people to reveal more accurately the sales potential of their districts. An sales person who believes that the sales potential of his district are great will take the income path that starts at a base salary of S_3. The sales person who does think the sales potential of his district are very good will choose the income path that starts at S_1.

What's a sales person to do? Lying about the sales potential of his or her territory won't help. Indeed, the sales person isn't even asked to lie. All he or she must do is choose from among the compensation packages in a way that he or she, not the manager, believes will maximize total pay. The sales person who sees little prospect for sales will choose the package with the salary of S_1. The sales person will be compensated for the limited sales potential by a high salary. The sales person who believes the sales potential will be greater than SP_1 (on the horizontal axis) but less than SP_2 will choose the package with a salary of S_2. The salesperson who believes that the "sky is the limit" (meaning a sales potential of greater than SP_2) will choose the package with the low salary of S_3.

This is the approach for establishing salary-commission rate pay contracts at IBM.[24] It's not a sure-fire way of making sales people totally honest, but it can help, and that's all real-world managers should strive to achieve.

Incentives in Rental Contracts

Because risk sharing and risk reducing contracts can be mutually beneficial, we should not expect two-part payment schemes to be restricted to payments by employers to workers. They can also be a part of the payments made by tenants to landlords. Rental agreements may not appear to involve "paying for performance," but surely they include performance pay. Both the landlord and tenant are intent on having an agreement that will ensure that the other will "perform" as specified. The landlord wants the rent. The tenant wants a nice living environment or, in the case of retail space, wants a profitable business environment. Each wants to get as much as possible from the other.

Consider the nature of rental payments within and near the city of Irvine, California, which is situated along the coast halfway between Los Angeles and San Diego. Irvine is a totally planned community with 110,000 residents and 140,000 jobs within an area of approximately 180,000 acres, or 42 square miles. It has been planned and developed not by the usual government planning boards, but by a private wealth-maximizing firm, the Irvine Company, which was once the Irvine Ranch.

One of the more interesting features of the city is that much of the commercial property continues to be owned and managed by the Irvine Company, which has an unusual contract with its commercial tenants. The contract requires that tenants make a three-part payment: a fixed monthly rental payment; a fixed monthly payment for upkeep of the common areas within the community shopping areas; and a payment based on a percentage of their profits. We are told that these payments can be quite stiff. For example, for a 1,000 square foot store in a shopping center called Fashion Island, an up-scale mall (actually in the adjoining city of Newport Beach), the rent can be several thousand dollars a month, plus several percentage points of the store's profits, plus several hundred dollars a month in maintenance fees, or so we have been told.[25]

How can the Irvine Company charge so much and then take a part of the store's profits? It is all too tempting to conclude, as many have, that the contract is "exploitive," reflecting the monopoly power of the Irvine Company. Maybe so. The owners and executives of the Irvine Company are wealthy. But, at the same time, there are good reasons to believe that the stores also benefit from the contract, especially a provision that gives the Irvine Company a stake in the profits of the stores in their shopping centers.

Naturally, any given store would love to retain the benefits of being in Fashion Island (or any other of the two dozen Irvine shopping centers) and, at the same time, pay

[24] This discussion of offering sales people a menu of contracts is taken from Paul Milgrom and John Roberts, Economics, Organization and Management (Englewood cliffs, N.J.: Prentice Hall, 1992), pp. 400-402.

[25] We are not privileged to the particulars of the contracts, but the exact dollars involved are irrelevant to our discussion.

no rent whatsoever. On reflection, however, the storeowner could easily see that such a deal would be a loser, unless it was virtually the only store that got such a deal. Each storeowner can reason that the payment for the upkeep of the grounds can clearly be in his or her best interests, given that the upkeep payments can make the whole center attractive to customers, increasing the traffic in all stores. These mandatory payments override the inclination of each storeowner to shirk on upkeep. The storeowners are, in effect, employing the Irvine Company to overcome the prisoners' dilemma problem they would otherwise face and that has been at the heart of so many other management problems considered to this point. They want the Irvine Company to perform with the interests of the stores in mind, as well as the interest of the Irvine company's stockholders. (Of course, there is a clear tie-in between the storeowners' interests and those of the Irvine stockholders. The better the storeowner's do, the better the Irvine Company does, and that's the kind of performance tie-in that the storeowners should seek.)

The storeowners can also reason that the high rental payments accomplish a couple of objectives. They ensure that all stores are high-value stores, with a focused appeal to up-scale shoppers. Low-valued stores are not likely to be able to meet the stiff rental payments. The high payments also ensure that prices will be somewhat higher at Fashion Island than at other shopping centers, thus causing downscale shoppers to go elsewhere (permitting up-scale shoppers freer access to the stores). The high rental payments also reflect the fact that the demand for the space at Fashion Island is high, and it is high simply because the Irvine Company has done a good job of enabling the storeowners to make high profits. Stores, in other words, don't always want low rents, because low rents usually go hand in hand with low profits.

But why would the stores ever *want* to sign a contract that enables the Irvine Company to share in its profits? Even this provision has an advantage for the merchants, given the conditions of the area. Storeowners understand that the Irvine Company controls much of the commercial space in the Fashion Island/Irvine area. The Irvine Company greatly influences the overall order of things in the area, including the income levels of residents, the distribution of various shopping centers within the Irvine area, and the distribution of stores within and across the shopping centers. The company has a terrific impact on the "look" and "feel" of the community, which means the company can greatly influence the degree of success of individual storeowners. (In many respects, the entire Irvine area can be viewed as one big shopping mall.)

Taken together, we should not be surprised that the Irvine Company takes a share of the stores' profits and that the store owners (collectively) *want* them to do just that. The percentage take gives the Irvine Company a direct incentive to operate in the interests of the storeowners. If the Irvine Company allows the community to deteriorate or allows "too many" direct (or even indirect) competitors into their shopping centers, then the company will suffer an income and wealth loss (given that the value of their shopping centers are a function of the stores' profitability). Hence, we would imagine that the standard contract is one that the storeowners like as much as the Irvine Company does, at least in terms of its basic features. The profit percentage is a way the storeowners can "pay for performance" on the Irvine Company's part.

We also should not be surprised that in many other areas of the country landlords do not include the percentage take. This is because in so many other areas, property ownership is often fragmented among a number of owners, with no one dominant property owner who is capable of determining, to a significant degree, the "look" and "feel" and profitability of individual store owners. As a consequence, storeowners are unlikely to give a percentage of their profits to their landlord when in fact the landlord can do little to earn the take. The landlord is unlikely to demand a percentage take because then the landlord would have to accept a lower fixed rental payment and would be at the mercy of the storeowner, who has complete control over the store's profitability. There are simply no mutual gains to be divided.

Put another, perhaps a more instructive, way, we should expect percentage takes to be a part of lease contracts where the landlords have a significant impact on store sales, for example, in shopping malls and other planned communities. The more fragmented the property, the less likely (or the lower) the percentage take.

We should only infrequently expect rents to be determined totally by a percentage take. The reason is the same as the one given above for the two-part performance pay system for workers: Both the landlords and tenants have an interest in sharing the risk. They both have an interest in a contract that reflects, to some degree, the influence that one party can have on the success of the other.

Spreading the Risk Costs

Business is full of risks, and it is full of risk sharing among owners, workers, suppliers, and even customers. Here, we have stressed that pay systems can be seen as a means by which employers and employees alike seek to share and spread the risk costs that are endemic to business. At its heart, the sharing will be a mutually beneficial exchange, with both parties accepting risk so long as the gains are greater than the risk costs incurred (or else the agreement will not last long). In addition, the pay system chosen is a means of inducing one party to act more effectively with the interest of the other party in mind. In a two-part pay system, the straight salary component (which can reduce the risk cost felt by the worker by more than his or her pay is cut) can encourage employers to ensure that there is work for employees. The piece-rate or commission component can encourage workers to work hard and smart.

How much should workers be paid in salary and commission? The answer is a disappointing, "It depends." The exact combination of pay components depends on such factors as the risk aversion of workers and how much the actual production levels in given work environments are under the control of workers and employers. The more risk averse workers are, the greater the salary component. This is because there is more profit to the firm by lowering its wage bill and accepting more risk of variations in worker incomes. The more output is dependent upon the actions of the workers, the greater the commission component. How should employers determine the combination? A good start would be for the employers to see if workers are willing to accept a reduction in their overall pay with more of their income from guaranteed hourly or monthly payments.

Of course, the firm will want to ensure that the reduction in compensation is greater than the added cost the firm calculates it will have to incur because of the reduction in production. The firm should continue to lower its overall wage bill that way so long as the reduction in overall pay is greater than the increase in the loss from slack output. It should, in other words, do what economists have long recommended -- "equate at the margins," balance the marginal gain with the marginal pain. In so doing, the firm cannot only achieve maximum profits, it can actually improve the welfare of its workers.

Most books in economics rarely, if ever, mention concepts like "integrity," "commitment," "credibility," or "bonding" in their discussions of how well the economy works. We give those concepts special attention because their importance far exceeds their notice. Managers depend on such basic notions. The competitiveness of firms depends on them. The efficiency of the economy depends on them. Managers and firms have failed simply because they did not give those concepts the respect they are due. Incentives tend to matter (in the right way) when, and to the extent, that managers' commitments matter.

MANAGER'S CORNER II: **Executive "Overpayments"**

Many workers at the bottom of the typical large corporate pyramid often grumble that their companies' executives are living off the fat of the workers' efforts, and that the executives could not possibly be worth their overblown annual salaries (often running into the tens of millions). On the surface, those who grumble seem to have a point.

The CEO of Time Warner, for example, made more than $137 million during the last five years of the 1980s, over half of which, $78 million, was received in one year alone (and $75 million of that year's compensation was in the form of a bonus provided as a reward for the merger of Time, Inc. and Warner Communications he helped orchestrate). However, a number of other CEOs made several tens of millions during the same period that the Time Warner CEO was pocketing his fortune.[26] The astronomical levels of executives' reported compensation prompt many workers and stockholders to argue that their companies would be better served if much of the executives' compensation and perks were used to pad the pay of lower-echelon workers.

At the same time, there is a less publicized trend in executive compensation packages -- CEOs who risk all, taking no salary, with their reward tied totally to the prices of their companies' stock through grants of stock options. When he was appointed CEO of Ingram Micro, Inc., a California-based computer distributor, in 1996, Jerre Stead took a wage of zero in spite of the fact that Ingram was reportedly ready to pay him $1.5 million in salary and bonus.[27] Stead insisted on having the right to buy up to 3.6 million shares of Ingram (or 2.8 percent of the company) over five years. Given that the company, at the time of Stead's appointment, was preparing to go public, Stead could be a wealthy man in spite of no wage. One compensation expert estimated, at the time, that

[26] See Steve Kichen and Eric Hardy, "Turnover at the Top," Forbes, May 27, 1991, pp. 214-218.
[27] As reported by Judith H. Dobrzynski, "Top Post at Rock-Bottom Wage: Chief Executive Puts Stock-Only Pay to Ultimate Test," New York Times, October 4, 1996, p. C1.

if Stead could triple the price of Ingram's stock over five years, his wealth could rise by a hefty $100 million -- nothing to sneeze at, to say the least.[28]

Of course, no one should forget that compensation tied to stock prices can translate into no gain and all losses (measured in foregone salary opportunities). When Nelson Peltz became CEO of Triarc Companies in 1993, a conglomerate in food, chemicals, and energy, he accepted an annual salary of $1 along with a bundle of stock options. As of 1996, Peltz had worked for nothing, given that he has been unable to exercise his stock options with the price of Triac stock falling by as much as 40 percent since the start of 1994.[29]

You can bet that some CEOs fiercely defend their high incomes, especially when their pay is dependent on their firms' performance. Former Scott Paper CEO Al Dunlap -- renowned for revitalizing dying companies with ruthless cuts in jobs, wages, and perks -- exudes pride for the $6.5 billion in additional wealth he made for Scott shareholders by radically downsizing and restructuring Scott: "My $100 million [in compensation, attributable in large measure to stock options he received and to additional stock he bought when he took over the head of the company] was less than 2 percent of the wealth I created for all Scott shareholders. Did I earn that? Damn right I did. I'm a superstar in my field, much like Michael Jordan in basketball and Bruce Springsteen in rock 'n' roll."[30] He adds that if there is criticism, it should be leveled against his predecessors at Scott who were running the company into the ground. His central admonition, all too easily forgotten, is, "You cannot overpay a good CEO and you can't underpay a bad one. . . If his compensation is not tied to the shareholders' returns, then everyone's playing a fool's game."[31]

The Tenuous Connection between Executive Pay and Performance

We agree that some workers have a complaint worthy of serious reflection. Many corporate leaders in this country are extraordinarily well paid and, we agree, some are probably "overpaid" (in a particular sense to be defined below), but not always for the reasons lower-level workers give. Even Dunlap acknowledges that "only a handful of chief executives are worth the big bucks they are paid. Many are grossly overpaid and should be fired and then replaced by CEOs whose pay is strictly performance-based."[32] Kenneth Mason, former president of Quaker Oaks, has much the same low opinion of the

---

[28] Ibid., p. C3.

[29] Dobrzynski, "Top Post at Rock-Bottom Wage," p. C1. However, according to compensation analyst Graef Crystal, Peltz's stock options had an estimated present value of $30 million or more (Ibid., p. C3).

[30] Al Dunlap and Bob Andelman, Mean Business: How I Save Bad Companies and Make Good Companies Great (New York: Times Books, 1996), p. 21.

[31] Ibid., p. 177.

[32] Dunlap and Andelman, Mean Business, p. 23. Dunlap doesn't mince many words when he adds, "In England, where I lived for three years, they have real royalty. In America, we have corporate elitists. Both are self-inflated windbags; they don't believe they're accountable to anyone. They enrich themselves at the expense of hardworking men and women who have actually invested in our companies. It's time they were accountable to someone" (Ibid., p. 209).

compensation packages received by many corporate heads, "It is a sad commentary on the intellectual vigor and financial discipline of the U.S. business community that so many corporate executives are receiving entrepreneurs' rewards for doing bureaucrats' jobs."[33] Moreover, it appears to be the case that the extent of executive "overpayment" is related to how much board members have invested in the companies they are asked to monitor: the greater the board members' financial stake in their companies, the less likely the executives will be "overpaid," with the converse equally true.[34]

However, it remains a safe bet that in many companies the higher up the corporate ladder the executive is, the greater the gap between his or her individual worth to the company and the pay received. However, it does not follow that the overpayments serve no useful purpose for the company or that any intentional policy of overpayment should be abandoned in favor of higher salaries for non-executives. Such a change in pay policy can have hidden perverse consequences for the company and its lower-level workers.

In making those points, and in showing the underlying logic below, we do not mean to suggest that companies do not make mistakes in executive compensation that should never be rectified. That, of course, would be a silly position to take. Business in all of its dimensions is filled with mistakes. We only mean to argue that there are good reasons for many corporate pay policies that result in the pay of executives exceeding their own individual marginal contributions to company income and profits.

Some of the high pay of executives is a reflection of intentional incentives included as a part of executives' pay contracts. Their pay is sometimes directly tied to corporate profits or to their companies' stock prices. Clearly, there have been cases in which executives' pay rises as firm profits sink and losses emerge. However, research shows a positive tie between company performance and executive pay. Indeed, finance Professor Sherwin Rosen found that top executive pay rises between 1 and 1.25 percent when the company's rate of return (as identified on the company's accounting statements) rises by 1 percent, not a bad deal for stockholders, given that most top executives' pay represents a minor fraction of company income.[35]

---

[33] Kenneth Mason, "Four Ways to Overpay Yourself Enough," Harvard Business Review (July-August 1988), p. 72.

[34] See Charles M. Elson, "Executive Overcompensation – A Board-Based Solution," Boston College Law Review (September 1993), pp. 937-996.

[35] Sherwin Rosen, "Contracts and the Market for Executives," (New York: National Bureau of Economic Research, working paper 3542, 1990). In the 1960s, economists speculated that large oligopolistic firms headed by managers who would be able to pursue their own objectives at the expense of stockholders would tend to base pay on sales, rather than profits or some other measure of direct stockholder wealth [see William J. Baumol, "On the Theory of Oligopoly," Economica, vol. 25 (August 1958), pp. 187-198; and "On the Theory of Expansion of the Firm," American Economic Review, vol. 52 (1962), pp. 1078-1087]. However, early researchers found profits, rather than sales, tended to govern executive pay [Wilbur G. Lewellen and Blaine Huntsman, "Managerial Pay and Corporate Performance, American Economic Review, vol. 60 (4, September 1970), pp. 710-72; and Robert Tempest Masson, "Executive Motivations, Earnings, and Consequent Equity Performance," Journal of Political Economy , vol. 79 (November 6, 1971), pp. 1278-1292). Both of these studies also found that firms that tied their executives' pay to firm performance also got better performance.

Much of the very high level of compensation is related to the fact that top executives are often given stock options, or the right to buy stock at a specified price, which means if the stock price goes up, the executive can do what everyone in the market wants to do, buy low and sell high. The executives' pay is as high as it is simply because their companies did well.

Now we understand, as critics of executive pay contend, that a firm's performance over time is dependent upon the actions of a number of people who are not always in the executive suite. However, we should expect executives to evaluate how much they contribute to the company, and their assessments should work into the pay deals that they demand. Executives who are considering the top position in a company and who believe a company will do well regardless of their contribution should be eager to work for that company, and the competition among the potential executive recruits should check the extent of the stock options and the price the executives will have to pay for the stock in the event the options are exercised. Competition will constrain the deals that are made. Many executives are extraordinarily well paid simply because their companies did far better than anyone could have expected when their pay deals were negotiated.

There is a good reason for concentrating pay incentives (especially those related to stock prices) on top level managers: they are the ones who control the most resources, whose decisions can have the greatest impact with firms, and who can be motivated by tying their pay to firm performance per se. Workers at the bottom of the corporate pyramid typically control few firm resources, and their *individual* actions (because each person is one of many similarly situated workers) are often immaterial to the performance of the entire firm. As a consequence, although we do not wish to be caught saying never, we stress that ties between pay of lower-level workers and firm performance may have little to no effect on the overall performance of the firm. This means that as pay incentives are extended down the corporate ladder, we should expect to see the extensions have progressively less impact on the performance of the company and, hence, the stock price, predictions that have been supported, albeit weakly, by empirical work.[36]

Admittedly, many firms do have profit sharing plans in which all workers share in the earnings of their companies. For example, Levi Strauss announced in 1996 a new incentive plan for all of its 37,500 employees that would reward, at the end of six years, workers with a bonus of as much as a year's pay if the company's profit goals were achieved. The plan could cost the company as much as $750 million in shared profits, but still the company must be betting that the incentive plan will increase profits by at least $750 million over what the profits would otherwise be in 2002.[37]

---

[36] The research found that announcements of incentive pay schemes (in the form of stock purchase plans) that were more inclusive than executives had lower effects on the price of the companies' stocks than did incentive pay schemes that were restricted to only the top or key executives [Senjai Bhagat, James A Brickley, and Ronald C. Lease, "Incentive Effects of Stock Purchase Plans," Journal of Financial Economics, vol. 14 (1985), pp. 195-215].

[37] As reported by Martha Groves and Stuart Silverstein, "Levi Strauss Offers Year's Pay as Incentive Bonus," Los Angeles Times, June 13, 1996, p. A1.

The fact that profit-sharing plans are available for many workers along with our logic outlined above suggests that we should revise our conclusion to the following: the lower down the corporate pyramid, the more tenuous or limited the connection between compensation and overall firm performance. The Levi Strauss incentive proposal may sound like a lot, but much less is at stake than might be initially thought, given that the worker will not receive a bonus for six years and, even then, the bonus may not match a full year's pay. If a full year's pay is paid at the end of six years, the annual bonus in present value terms can average less than 10 percent of a worker's annual pay over the next six years.[38] If the bonus is further discounted by the probability that not all of it will be received (due to resignation or firing), the expected value of the bonus can easily be a minor fraction of the salary.

Providing workers with stock options or even shares of stock is a way of giving them a stake in their companies and an incentive to do that which the owners want them to do: work hard to increase firm profits and, thus, the price of the stock. If the firm's stock price goes up, the workers can gain by exercising their stock options (buying at the stipulated price of, say, $10, and selling at the going market price of, say, $22) or just selling the stock for $22 (which they may have earlier been granted instead of a wage increase of $10). However, while the practice of giving workers some stake in their firms through shares of stock appears to have been growing in the 1990s, it is still not widespread among major U.S. companies. Only about 3 percent of the top 1,000 U.S. corporations granted *all* workers some stock stake, either in the form of options or outright shares. Between 8.5 percent and 13 percent provided a stock stake to more than 60 percent of their employees.[39]

Executive income can be far more dependent upon built-in incentives. However, it does seem reasonable to conclude that if strong incentive pay for executives has its intended effect, lower-level workers can also be better off than they would have been otherwise, given that their incomes and job security are enhanced by executive decisions that lead to higher profits and stock prices. Lower-level workers, in other words, can have an interest in seeing their bosses' incomes, but not necessarily their own, strongly tied to firm performance. And the evidence does suggest that when the pay (salary plus bonuses) is evaluated across firms with varying rates of return on common stock through time, a positive relationship is evident: the higher the rates of return, the higher the executive pay. In addition, executive *total* compensation (including salary, bonuses, and benefits from stock options and stock grants) appears to be strongly related to firm

[38] A worker who this year is paid $25,000 and is expected to be paid $30,000 in six years (assuming a cost of living raise of 3 percent a year) will receive a bonus of $30,000 in 2002, assuming the firm's profit goals are reached. The present value of the $30,000 bonus is, however, only worth slightly more than $15,000 today (assuming an interest rate of 12 percent). The bonus will amount to less than 10 percent of the worker's annual income.

[39] As reported by Michael A. Hiltzik, "More Firms Giving a Stake to Employees," Los Angeles Times, June 15, 1996, p. 32, based on a report from the Executive Compensation Reports.

performance. The greater the firm performance, the greater the total compensation of executives.[40]

Economics Professor Kevin Murphy found that for executives who worked for companies that in the 1974-1979 period had *negative* rates of return greater than 20 percent, the average annual change in executive pay was a mere half percent.[41] Murphy also found that pay increased with greater rates of return, reaching a nearly 11 percent increase in pay for those executives whose companies had positive rates of return greater than 40 percent. In the later 1980-1984 period, the pay of the executives working for companies with rates of return of greater than 40 percent increased by 17 percent.[42]

Furthermore, Murphy found that the changes in the prices of the executives' stock holdings could dwarf the changes in their compensation (or even their absolute levels). Executives who worked for companies with greater than *negative* 20 percent stock price returns suffered an annual average decline in the value of their stock holdings of nearly $3 million (at the same time that their pay averaged $506,700). Those who worked for companies with a greater than positive 40 percent stock return realized an increase in the value of their stock holdings of $3.7 million (at the same time that their average pay was $494,300).[43]

Professors Michael Jensen and Kevin Murphy found that every $1,000 increase in stockholder wealth corresponds to just over 2 cents more in CEO median annual cash pay but a $3.25 increase in median executive wealth,[44] a finding that caused one of the authors to conclude in the Harvard Business Review that "top executives are worth every nickel they get."[45] Critics, however, may rightfully charge that top executive pay is not sufficiently dependent on firm performance and should be dramatically raised, as Kenneth Mason has charged.[46] The relatively weak connection between the fortunes of executives and stockholders may be explained by the fact that CEOs can be easily

---

[40] However, bonuses appear to be more strongly related to management performance than are merit increases [Lawrence M. Kahn and Peter D. Sherer, "Contingent Pay and Managerial Performance," Industrial and Labor Relations Review, vol. 43 (special issue, February 1990), pp. 107s-120s].

[41] Kevin J. Murphy, "Top Executives Are Worth Every Nickel They Get," Harvard Business Review, March-April 1986, pp. 125-132.

[42] Ibid., exhibit I, p. 126.

[43] Ibid., exhibit III, p. 129.

[44] Michael C. Jensen and Kevin J. Murphy, "Performance Pay and Top-Management Incentives," Journal of Political Economy vol. 98, no. 2 (1990), pp. 225-263. The tie between stockholder wealth increase and the increases for the executives varies by the market value of the firms. Executives who headed the firms in the "bottom half" of the firms studied, measured in terms of firm market value, had a median increase in personal wealth of $8.05 per $1,000 increase in stockholder wealth. Those firms in the "top half" had a median increase of $1.85 per $1,000 increase in stockholder wealth. Other studies have found stronger ties (perhaps eight times stronger) between executive compensation and firm performance. See Peter F. Kostiuk, "Executive Compensation, Corporate Performance and Managerial Income," Center for Naval Analysis, January 1986, who found that executive compensation rose by 12.5 percent when the accounting rate of return rose by 10 percent, and Andrew Cosh ["The Remuneration of Chief Executives in the United Kingdom," Economic Journal, vol. 85 (no. 1, 1975), pp. 75-94], who found that executive compensation rose by 10 percent when the accounting rate of return rose by 10 percent.

[45] Murphy, "Top Executives Are Worth Every Nickel They Get."

[46] Mason, "Four Ways to Overpay Yourself Enough," p. 73.

monitored, evaluated, and dismissed by their board members, but firings appear to be used very sparingly as a means of discipline.[47]

Jensen and Murphy suggest that the weak connection between CEO compensation and firm performance and the very limited use of firings may be attributable to the fact that public disapproval (and attendant political considerations) of high salaries may impose artificially low upper bounds on CEO compensation. Hence, in order to attract CEOs (by ensuring that the expected value of the compensation package is maintained), the boards have to limit pay cuts and, for that matter, firings.[48] Still, the market appears to believe in the future value of current announcements of executive pay plans that tie the executives' long-term compensation to the long-term performance of firms through outright stock grants, stock options, and bonuses. According to one team of researchers, firms that install incentive plans for their executives can expect to see their stock prices jump by 2.4 percent within two months over and above what they would otherwise have been.[49]

Surely, however, direct incentives for executives do not explain all of the sometimes-exorbitant levels and growth of some executives' pay, nor would we expect explain everything about executive pay.[50] Stockholders and their boards must be concerned with incentives for lower-level workers as well when they set executive compensation levels. As noted, executive compensation can be used to give aspiring executives within the firm an incentive to work hard.

Granted, the high pay of executives can be *partially* explained by the fact that the people who become executives generally get their positions because they have demonstrated that they are more capable than other workers. Moreover, a move to a higher-ranking position can actually increase the productivity of the manager. As Rosen has observed, "Scarce talents of the most capable managers are economized by assigning them to positions at or near the top of the largest firms, where their ability is magnified to greater effect by spreading it over longer chains-of-command and larger scales of operations. This is what sustains high average earnings of top level executives in large

---

[47] CEOs whose rates of return match industry standards have only a 4 percent chance of relinquishing their jobs, according to Jensen and Murphy. CEOs whose rates of return are 50 percent below industry averages have a three times greater chance of relinquishing their jobs, but still the probability is only 12 percent and then the turnover may be voluntary, due, for example, to retirement (Jensen and Murphy, "Performance Pay and Top-Management Incentives," p. 20).

[48] In contrast to the claims of critics of executive compensation, Jensen and Murphy have found that CEO compensation actually declined in real dollar terms between the 1930s and 1980s as firm values increased. The incentive executives have to work in their stockholders' interest has also declined, given that the wealth gains to the executives per $1,000 of stockholder gains has declined (Ibid., pp. 253-260).

[49] See James A. Brickley, Sanjai Bhagat, and Ronald C. Lease, "The Impact of Long-Range Managerial Compensation Plans on Shareholder Wealth," Journal of Accounting and Economics, vol. 7 (1985), pp. 115-129.

[50] One explanation for the perceived growth in executive compensation in the 1980s is the method of reporting executive pay. Prior to 1978 firms could place executive compensation in the form of stock and stock options at the back of their annual reports, where such pay factors could go unnoticed and unreported in the media. In 1978, the Securities and Exchange Commission began requiring firms to put all forms of executive compensation in the front of the annual report where investors and reporters could more easily notice them.

firms and also implies that firm size and executive pay should be positively related," which has been shown to be a pervasive feature of executive pay.[51] Hence, they not only deserve higher salaries, they must be paid higher salaries because, if they are not, other firms will hire them away.

Once someone is promoted to the executive ranks, his or her pay must also go up significantly at the time of the promotion simply because the executive becomes more visible to the rest of the relevant business community. Before the promotion, other firms might be unaware of the executive's abilities. After all, he or she might be toiling away with a team of other workers where his or her abilities can be difficult to evaluate, especially by outsiders. By promoting a person, a company announces to other firms that they have found someone in their midst who is unusually productive and might even be on a fast track to the top office in the firm. Outsiders no longer have to incur the costs associated with searching through a large group of some other firm's workers to find productive managerial talent. They can "cherry pick," limiting their picking to the "cherries" identified by others.

The gap, which can be substantial, between the pay of those who are promoted and those under them can be partially explained not so much by their actual productivity as by the fact that the more productive workers at the bottom of the corporate ladder have not yet been "discovered," and, just as in the case of aspiring actors, managers understand -- or should understand -- that being "discovered" can be as important in rising through the ranks as actually acquiring the skills to undertake higher level jobs. Not all people with the acquired skills (many of whom may be reading his book) will make it onto the upper rungs of the corporate ladder.

Hence, outsiders can be expected to target those who are promoted elsewhere, competing with the newfound executive's own firm. Put another way, a firm must make promotions count in terms of added pay and all the trappings that can go with higher office as a defense against "executive raiders" intent on minimizing their search costs for managerial talent.

Rising through the ranks probably requires a dose of luck and political acumen, with both considerations having little to do with actual productivity, as many people would measure it. Many workers no doubt grumble about executive pay with cause. They, the grumblers left behind, may in fact be more productive than some of the people above them; they just haven't met with the requisite measure of luck. Also, being discovered often requires work at getting oneself noticed through, for example, self-promotion, and the time devoted to such activities can be time taken away from improving one's managerial skills. Moving up the ladder on the fast track requires not just managerial skills per se, it requires some *optimum* combination of skills and self-promotion and schmoozing. There are no doubt many workers left behind who are indeed more productive than those who are promoted; they just never found the right use of their time. In effect, they have acquired "too much" in the way of basic skills and not enough of, say, political savvy.

[51] Sherwin Rosen, Contracts and the Market for Executives (New York: National Bureau of Economic Research, Inc., working paper 3542, December 1990), p. 7.

Just because pay differences between the ranks may be partially based on luck, it does not follow that the differentials should be eliminated, even if they could, which they probably could not be, given competitive forces. All corporations can be expected to do is establish promotion and pay policies that will enable them to achieve a reasonable measure of success -- not perfection -- in picking the "best" people for higher level jobs. If they sought perfection in the selection process, the companies would surely fail simply because mistakes are usually unavoidable in most complex business/employment environments. In their quest for perfection, the companies would also incur excessive search costs, making them uncompetitive vis a vis other companies that were willing to accept occasional mistakes.

Executive Pay As a Motivation for Workers

The pay of executives may also be "excessive" for another reason involving the difficulties of selecting managers. When people are hired at the bottom of the corporate ladder, upper level managers may have only a rough idea as to whom among the large group at the bottom are worthy of higher ranks. They can, for example, check references and look at their workers' educational records -- what schools they attended and what grades they made -- but such factors are not always highly correlated with a willingness on the part of people to work hard and smart in given corporate environments.

How can upper-level managers motivate lower-level workers to reveal how hard and smart they are, at the limit, willing to work? Piece-rate pay and two-part pay contracts, which we have covered, can help. So can bonuses. Another incentive system used is an executive "tournament," which is held among lower-level workers, with the "prize" being a promotion to the next rung on the corporate ladder.

Any overt or covert announcement of the tournament can have two effects. First, it can cause the workers to compete among themselves for the prize. All workers can work harder for the prize with the added value being claimed by upper managers and owners who announce the competition.[52] Second, aware of the competition among employees, workers who might be hired at the lower levels in the firm with the tournament will self-select. Those who think that they will not "win," and who will therefore suffer the cost of the competition but will not receive a "prize," will self-select out of employment with the firm.

Therefore, the tournament will tend to be concentrated among those who have a degree of confidence in their abilities, given the competition. Workers who self-select

[52] The executive tournament can have much the same effect as prizes do in real golf tournaments: they improve performance. One study found that by raising the prize money to a hundred grand or more, the scores of the golfers went down by 1.1 strokes over the course of a 72-hole tournament. Apparently, the prize money had its greatest effect in the later rounds when the players were tired and needed to concentrate on every shot [Ronald G. Ehrenberg and Michael L. Bognanno, "Do Tournaments Have Incentive Effects?" Journal of Political Economy, vol. 98 (December 1990), pp. 1307-1324]. In addition, bonuses appear to be sensitive to managerial bonuses with the future performance of managers improving with current bonuses [Lawrence M. Kahn and Peter D. Sherer, "Contingent Pay and Managerial Performance," Industrial and Labor Relations Review, vol. 43 (February 1990), pp. 107S-120S)].

into the competition can then compete in the knowledge that their cohorts at work will, on average, be more productive than they would have been if the tournament were not held. Their *expected* lifetime pay with the firm should, accordingly, mirror the higher expected productivity of the workers hired.

In order for the tournament to have the intended effect, the pay upon promotion (or winning) must be attractive to all who compete at the lower levels -- after the higher pay is discounted by the probability that any one person will receive it. In group settings, most reasonable worker/competitors will likely assume that the probability of their being selected for the promotion is significantly below 1.0 (or certainty). After all, when they start the contest, the competitors will have only limited information on just how hard and smart their cohorts will apply themselves. And pay and the probability of promotion do appear to be inversely related. According to one study, pay increments with promotions increase substantially between managers at adjacent levels within corporations, and the pay increments when promoted vary inversely with the prospects of being promoted, which should be expected: the stiffer the competition (and the lower the prospects of being promoted), the greater the pay increase must be in order to maintain the drive among managers to be promoted.

Those participating in tournaments should demand a higher expected pay because tournaments are by nature "games," meaning the outcome is dependent upon how the other participants play, or seek the prize. This aspect of tournaments necessarily introduces some variance in the outcomes of tournaments, which implies unavoidable uncertainty into how individual participants should "play" (or compete). The pay should be expected to compensate the participants for the problems associated with the inherent risk and uncertainty (vis a vis other pay systems – for example, piece rate – that simply require the workers to maximize their output without consideration to what other workers do).[53]

Therefore, the value of the prize (which includes an "overpayment") must be some multiple of the total costs each worker can be expected to expend in seeking the promotion. The lower the probability of any one worker receiving the prize, the greater must be the value of the prize -- the overpayment, or the gap between the promoted person's actual worth to the company and the pay (plus fringes and perks). If the gap were nonexistent, then the prospects of promotion would not have the intended impact a tournament is supposed to have on all workers' productivity. [54]

---

[53] For a discussion of these points and some experimental evidence that suggests that the variance of outcomes in tournaments is greater than the variance in outcomes of piece-rate pay systems, see Clive Bull, Andrew Schotter and Keith Weigelt, "Tournaments and Piece Rates: An Experimental Study," Journal of Political Economy, vol. 95 (no. 1, 1987), pp. 1-33.

[54] See Jonathan S. Leonard, "Executive Pay and Firm Performance," Industrial and Labor Relations Review, vol. 43 (no. 3, 1990), pp. 13s-29s. Also, consistent with the Leonard study, another study found that pay increases rapidly with higher ranks, with the CEO earning $100,000 more a year than vice presidents compared to lower-level managers earning $10,000 to $30,000 more than their underlings [Richard A. Lambert, David F. Larcker, and Keith Weigelt, "The Structure of Organizational Incentives," Administrative Science Quarterly, vol. 38, no. 3 (September 1993), pp. 438-462. However, another study drew a contradictory conclusion: that the greater the number of vice presidents (which, presumably means a lower probability of being promoted), the greater the pay gap between the CEO and the vice presidents [C.

Put another way, promoted workers usually get substantial pay increases with larger offices and more perks not because they necessarily "deserve" all that they get, but because the firm may want to validate the tournament and to hold other tournaments in the future. The executive's "overpayment" is covered by the firm not so much by what the chosen executive actually does (although, as noted, that can be an important factor), but by the added output generated by the competition among all those who seek promotion.

Why is it that pay rises so fast as people are promoted through the ranks? Again, there is, no doubt, some correlation between rank and abilities, although it is by no means perfect. The higher up the ladder, the greater the abilities of executives -- as a tendency. However, we suspect that pay differences have a lot to do with probabilities. Someone at the bottom looking up the ladder can figure that the probability of his or her actually making it through the rungs falls the further up the ladder he or she looks. A worker at the bottom might give him or herself a probability of 20 percent of making it to the first rung, given the few people in the immediate work group, but the worker might give himself or herself a probability of .001 percent of making it to the top rung (and even that probability might be overstating the prospects of success), given that he or she might be competing with everyone in the organization and those who may join the organization in the future. And the worker is likely to reason that the greater the number of workers at the bottom and the greater the number of rungs in the corporate ladder, the smaller the probability of reaching the top rung.

Executive pay, in other words, must rise disproportionate to productivity just to account for the declining probability of any one person making it through the rungs. The purpose of the progressively larger "overpayments" at the higher and higher rungs is not necessarily so much designed to promote social justice among workers, although such considerations are rarely totally overlooked either, but it is to properly motivate all workers who are contemplating moving through the corporation.

The Growing Gap between Executive and Worker Pay

Again, why is it that the pay gap between top executives and workers at the bottom has been growing over the last decade or so? Popular wisdom has it that the growing gap can be attributable to insane corporate policies that are stacked in favor of executives by board members who were appointed to their positions to do what they have done, raise the income of the executives at the expense of owners and lower-order workers. According to Graef Crystal, a prominent critic of corporate pay, boards of directors not only raised their CEO pay by an average of 21 percent in 1995 (several times the rate of inflation), but they raised pay for reasons that are hard to identify. Ten percent of the variation of pay among top executives can be explained by company performance: better performing companies tend to pay their CEOs better. Twice that percentage (21 percent)

O'Reilly, Brian Main, and G. Crystal, "CEO Compensation as Tournament and Social Comparison: A Tale of Two Theories," Administrative Science Quarterly, vol. 33 (no. 3, 1988), pp. 257-274.

of the variation can be explained by company size: larger firms tend to pay their CEOs better. That leaves 69 percent of the variation unexplained.[55]

There is always a hint of truth in such claims, but we aren't willing to concede that none of the unexplained variation (just because it isn't picked up in regression analysis) in corporate pay has a rational basis. Corporate boards do some pretty stupid things from time to time (which market pressures force them to correct or suffer the consequences). However, we suspect the growing gap has something to do with the actual impact of executives on corporate earnings, given their decisions can be more important in a rapidly changing global economy, and with the declining opportunities of workers making it to the executive suite, given the "flattening" of corporate command-and-control organizational structures. The probability of someone becoming a chief executive officer has simply gone down at the same time that the risk of being an executive has gone up.

We should also not overlook the prospect that the high pay of the top executives in a firm may be a means of driving down the pay of the workers at the bottom. Indeed, that can be the purpose of the overpayment of the people at the top. By raising the pay of executives, more people can be attracted to the firm in the hope that they will eventually make it to the top and receive the overpayments. In this sense, there is not only a gap between higher and lower worker pay, there is also a gap between what the lower workers are paid and their expected pay, and the gap between the actual and expected pay of lower workers can expand as the gap between the actual pay of the lower and higher workers increases.

All of this means that workers may indeed be right when they complain that their chief executive could not possibly be worth the zillions that he or she makes. "Worth" is not necessarily the point of the pay. Properly aligning the incentives of workers throughout the organization is the point that should not be overlooked.[56]

The overpayments provided executives can, of course, be fortified by market competition for executive talent. All firms interested in maintaining proper incentives can compete with each other for executive talent, but their competition can be constrained by the fact that they cannot wipe out their overpayments. If they did, then incentives, and production, throughout their firms could be impaired.

---

[55] Graef Crystal, "Average U.S. CEO Boosted Pay 21% in '95, to $4.5 Million," Los Angeles Times, May 26, 1996, p. D4.

[56] We don't want to be accused of playing to the view that executives are the only group of workers who can be "overpaid." We presented arguments much earlier in the book as to why some workers are "overpaid." Obviously, in many firms there are also workers who become good at working the pay system to their advantage without their bosses noticing. They can end up overpaid for a very long time. Also we are sympathetic to the view that many executives are probably "underpaid," given how little their rewards go up with their executive actions. At the same time, many workers may be overpaid, given how little they can affect their company's revenues for the wages they receive. A contrarian view is developed at length by Robert H. Frank, Choosing the Right Pond: Human Behavior and the Quest for Status (New York: Oxford University Press, 1987).

Executives can also be "overpaid" because they are in positions of trust, and they have command over large amounts of firm resources. Typically, the higher up the executive, the greater the resources that the executives can direct. Firms want to make sure that the executives do not violate their fiduciary responsibilities. One method of discouraging violations is to ensure that the executives incur a significant cost if they are ever fired, and that objective can be accomplished partially by paying executives more than they are "worth" in the market. Hence, we can conclude that the overpayment will be related to the probability of executives' misdeeds being detected as well as the damage that the executive can do to the company if he or she ever succumbs to the temptation to violate his or her responsibilities.

In general, the lower the probability of detection, the greater the need for a penalty -- and pay premium; and the greater the damage that the executive can do, the greater the pay premium.

Overall, what the stockholders want to do is align the private interests of their chief agents -- the executives -- with their own interest, which is maximizing the value of their investment portfolios, and stockholder portfolios can include shares in a variety of companies. As we have noted before, stockholders may naturally be less risk averse than their executives who can have a high percentage of their own personal portfolios -- including their human (managerial) capital -- tied up in the firms they manage. Executives may understandably worry about the failures of their particular companies, which can undercut the market value of their human capital. Therefore shareholders are better off when executives face incentives that reduce their reluctance to take risks.

Stock options are a means of eliminating some of the downside risks managers face. The executives gain only if the stock price rises and do not lose if it falls. Often, the high levels of executive compensation reflect the exercise of stock options, which were made a part of their contracts simply as a means of encouraging them to take calculated market risks that their bosses, the stockholders, want them to take.

That is to say, executives may be the highest paid workers in a firm because more of their pay tends to be at risk; they need extra compensation for accepting the extra risk. And stockholders want it to be that way, given the considerable discretion top executives have and the influence they can have over firm performance. Lower ranking managers will not have as much discretion, nor will they likely have as much influence over firm performance. Their bosses will largely check their actions. Hence, lower ranking managers can be expected to have a smaller share of their pay at risk, leading to a smaller risk premium than the top executive receives.

Now, we don't want to overlook the fact that executives, like lower-level workers, can shirk their responsibilities, and engage in opportunism, one form of which is using the powers of their office to appoint board members who are willing to go along with pay increases for the executives. This form of overpayment can be disparaged for many reasons, but it remains a reflection of the principle/agency problem that has been at the heart of most topics in this book. Such "overpayments" may, in some sense, be "wrong," but we are not so sure that anything can or should be done about *all* such overpayments. Eliminating all such forms of opportunism is simply impossible, and the best

stockholders and boards can be expected to do is to minimize this source of overpayment. All we can say is that we should expect that the more difficult it is to monitor executives, the more likely they will be overpaid, or the greater the overpayment.

Needed Stability of Executive Pay

Of course, executive compensation as a *process* is far more complicated than simply that of setting a compensation package for executives that is, for example, heavily weighted toward rewarding executives for their companies' performance, whether measured by the bottom line or stock prices. It may be a great idea, for example, to tie compensation to stock prices. Executives will like that -- so long as they expect the price of the stock to rise. The problem is not the concept, but with application of the concept in practice. Any compensation scheme that is installed can be uninstalled, and executives can be expected to work for a change in their pay-for-long-term-performance scheme if their stock prices start going down. To the extent that the compensation scheme is changed (or can be changed), it can lose much of is potential incentive benefits. Executives can figure that they need not press for performance because they can, at some future point, shift their compensation from stock to salary. (The problem of adjustments in executive pay is hardly trivial, given that one study in the 1970s and 1980s found that the compensation incentive plans in the country's 200 largest industrial companies had an average life of 18 months.[57]) Moreover, stockholders may not want to *always* hold firmly to their pay-for-long-term-performance pay scheme, given that they may begin to lose valuable executive talent with downturns in the prices of their stock. This is especially true if stock prices fall because economic conditions beyond the control of the executives turn against the company.

Therein lies an applicable principle: compensation schemes should have some rigidity and should be changed only when firm performance cannot be attributed to management. It goes without saying that the more control executives have over their own compensation, the less effective will be any set of incentive plans. Then again, any rule that allows payment adjustments attributable to forces external to the firm leaves open the prospects for executive opportunism; executives can claim that firm performance is "someone else's fault." Therein lies an even more basic principle: boards of directors and their appointed compensation committees must be willing to stand tough. There's simply no escaping the need for tough judgments in business. Otherwise, the firm will risk being a takeover target.

Huge Exit Pay for Executives

There is an emerging trend in executive compensation that often rankles even some of the more staunch defenders of high executive pay: the growing tendency of firms to provide their executives with huge payoffs when their firms fail and/or the executives are fired.

[57] The study covered from 1975 through 1983 (as reported by "Four Ways to Overpay Yourself Enough," p. 71).

John Walters, whom AT&T employed as president with an eye toward later making him CEO, was granted a payoff of nearly $26 million after the board reneged on its agreement to promote him. The board members concluded that he was not up to the job he was hired to do. Michael Ovitz walked out Disney's door after only 14 months on the job with a $90-million payoff, while Gilbert Amelio left Apple Computers after only 17 months with a $7 million payoff.[58]

How can such payoffs be justified, if at all? Maybe the payoffs are a form of board graft, which is often implied when the payoffs are mentioned in the media. If that were all there was to it, it would appear to us that the firm that systematically did such things would be a takeover target.

Clearly, we suspect that there is more to the matter than greed and graft, although we don't want to totally dismiss such concerns. People and firms are imperfect, which is a theme underlying most economics discussions. We simply note that the payoffs can provide benefits for the company, mainly in the form of avoiding costly suits from fired executives. The payoffs may be "high," but still "lower" than the realistic options. The payoffs also enable the company to move swiftly –that is, to move failed executives out the door with a view toward replacing them with talented people who can do a better job. The firms can avoid the considerable damage an executive could do – through action or inaction -- to the firm if the payments are not made and the executive lingers in the job for months while the board attempts to negotiate a more modest payoff.

But, often the payoffs are nothing more than payments that fulfill the terms of the executive's contract with the firm. Knowing that they can be fired in short order at the will of the board, smart executives have negotiated the dismissal payoffs. The payoffs are simply the "tit" in "tit for tat" deals. In making their employment deals, firms must realize that they will invariably be seeking to pull an executive away from a known employment circumstance, which may carry with it substantial security because of the record the executives might have established, and place the executives in a less well known and, therefore, more insecure employment circumstances. The firms can expect to pay, in one way or another, for the added insecurity the firm effectively asks the executives to assume (and the greater the insecurity or risk of being fired, the greater the added payment, a force that will cause firms to pause in their willingness to act recklessly). Also, in agreeing to the new employment deals with dismissal rewards, the executives have, in effect, possibly given up something in the way of the level of their compensation, if they are able to stay with the firm, for the security that comes with the dismissal payoffs. The firm also benefits in such a deal, given that they know what the limits of the payoff will be, in the event the firm elects to fire the executive. Presumably, the bargain is expected to be mutually beneficial to both the executive and firm.

Granted, firms often make mistakes; they end up agreeing to pay deals for executives who prove to be "losers," but firms are in the business of taking such risks. The contract with any given executive can be seen as nothing more than a risky investment (or business venture) among an array of similarly risky investments (or

[58] See Judith H. Dobrzynski, "Growing Trend: Giant Payoffs for Executives Who Fail Big," New York Times, July 21, 1997, p. A1 and A10.

ventures). This means that executive payoffs must be judged not by how they work in individual cases of miserable failures involving outlandish payoffs, but in terms of how the "portfolio" of such deals payoff in the aggregate. This is to say that AT&T and Disney, and their stockholders, may have lost handsomely in the cases the fired executives already cited. However, the *approach* they have taken could be working very profitably, a fact that is often not mentioned in news reports of the lavish payoffs firms provide their failed executives.

There is another justification for the executive payoffs that seeks to overcome the different circumstances of the executives and stockholders. Members of the board can understand that executives might be more reluctant to pursue risky ventures that offer the prospects of high returns than the stockholders. After all, the stockholders can have highly diversified investment portfolios, with shares owned in a number of companies (or mutual funds). The stockholders also do not have their human capital invested in the firms they own. The executives are indeed different. By taking the jobs that they do, they invest their human capital in a given firm, and they put their human capital at risk. Because of the extent to which their compensation package may be heavily weighted toward stock and stock options in their firm, the executives can easily have a portfolio that is less diversified than the firm's stockholders. The lack of diversification can be an important pressure on the executive to "play it safe." The executives can lose their careers with risky investments; as we have seen, they may not gain nearly as much as their stockholders/residual claimants in the event that risky investments actually pay.

The dismissal payoffs for executives can simply be a means by which firms can encourage executives to take more risk, and thereby more closely align executive interests with stockholder interests. With the guaranteed payoffs, the firms are saying to their executives, "If you fail, some of your loss will be covered. Hence, we encourage you to take risks." The payoffs can also send a message to executives that are contemplating taking the top jobs, "If you fail, you will also be covered, at least in part." Accordingly, firms that do not make the payoffs on dismissal can be hiking their costs of recruiting executives and/or may have to settle for less qualified executives.

Firm Size and Executive Pay

Research shows that executive pay rises with the size of firms. The larger the firm, the greater the executive pay. According to one study of executive pay at 73 large corporations in the United States between 1969 and 1981, a firm with 10 percent more sales will, on average, pay their executives 2 to 2.5 percent more in annual salary plus bonus, an estimate remarkably close to the sales-pay relationship found by the researcher for the 1937-1939 and 1967-1971 periods.[59] Other studies on executive pay in the United States and Great Britain have found similar ties of executive pay to firm assets, that is, when firm assets grow by 10 percent, executive compensation grows by 2.5 percent to

[59]Peter F. Kostiuk, "Firm Size and Executive Compensation," Journal of Human Resources, vol. 25 (no. 1, 1989), pp. 90-105. See also Kevin J. Murphy, "Corporate Performance and Managerial Performance," Journal of Accounting and Economics, vol. 7 (no. 2, 1985), pp. 11-42.

3.2 percent (which may explain why executives often seek to expand into areas that have nothing to do with their core line of business, which may dampen profits, but raise executives' pay).[60]

We frankly don't know whether these findings are "good" or "bad" for the firms involved. On the one hand, the rise in pay may reflect the rise in the ability of executives to engage in opportunism, but, as stressed, it may also reflect a growth in the actual productivity of executives as they move up the corporate ladder. The more productive managers are, the more likely they are to be promoted, and any move up the ladder will necessarily increase the manager's productivity simply because his or her actions will radiate down the corporate hierarchy through more people.[61] On the other hand, the rise in pay may reflect an intentional policy to encourage lower workers to work harder and smarter. As firms grow, they need higher pay for executives in order to enhance incentives and get more production from workers down the hierarchy (or to offset the tendency of workers down the hierarchy to shirk as the firm expands).

All we can really say in closing is that high executive compensation often times makes more economic sense than commentaries in the popular press would lead readers to believe. Stockholders, board members, and upper management need at least to think about how they can manipulate their executive pay structure, up and down the hierarchy, as a means of making money for their firms. Higher executive pay can mean more work and output from people who have not yet been chosen for the executive suite, and most of whom will never be chosen (although many will make every effort to be chosen).

At the same time, the executives themselves must be mindful of the fact that market forces are also afoot that can ultimately check what they can do and how much they are paid. Executives whose companies do poorly because of their misguided decisions and opportunism can anticipate that their market value will suffer with a drop in

---

[60] See Cosh, "The Remuneration of Chief Executives in the United Kingdom," Economic Journal, vol. 85 (no. 1, 1975), pp. 75-94; Jason R. Barro and Robert J. Barro, "Pay, Performance and Turnover of Bank CEOs," Journal of Labor Economics, vol. 8, no. 4 (October 1990), pp. 448-481; and Joseph W. McGuire, John S.Y. Chiu, and Alvar O. Elbing, "Executive Incomes, Sales and Profits" American Economic Review, vol. 52 (no. 4, 1962), pp. 753-761.

[61] This theory can explain why one study found that managers located at their corporate headquarters tended to receive greater bonuses for performance than did their counterparts located away from the headquarters. The managers at the headquarters can potentially have a greater impact on more people and, accordingly, are potentially more productive (Kahn and Sherer, "Contingent Pay and Managerial Performance," pp. 107s-120s).

offers. Their firms may also be subject to takeover, given that bright investors can buy the firm, replace the existing management team with a more competent team, and then sell the firm at a higher price. The poor performance of one management team can represent a profitable opportunity for their competitors in the market for firms and management talent.

Concluding Comments

In a competitive labor market, wage rates are determined by the interaction of willing suppliers of labor (employees) and demanders of labor (employers). Suppliers are influenced significantly by the nonmonetary benefits of employment, as well as by the value they place on their next-best alternative employment. Thus differences in money wage rates may not reflect true differences in full wage rates. Demand is influenced by the laborer's productivity and the price of the laborer's product.

In a competitive labor market, any attempt to change workers' incomes through minimum standards for wages or working conditions can benefit some workers only at the expense of others. As the economist Milton Friedman complains, "The old saying is that Quakers went to the New World to do good and ended up doing well. Today, well-meaning reformers go to Washington to do good and end up doing harm."[62] Can the mixed results of minimum wage legislation be construed as a clear-cut improvement in social welfare? Economic analysis cannot address that highly subjective question. The best we can do is present the deductions drawn from theory and evidence. Unfortunately, both are conflicting, as evident in the theoretical implications of minimum-wage hikes under competitive and monopsonistic conditions and as evident in the differing empirical findings.

We have also shown that while it is nice to suggest that workers be paid according to performance, the issues of providing the "right" pay for the "right" performance are thorny ones for managers. Regrettably, the reality of managing can be tricky, as evident in our discussion of executive pay, or, rather, "excessive" pay. There are good economic explanations for executives to be paid more than they are "worth."

Review Questions

1 The government requires employers to pay time-and-a-half for labor in excess of forty hours a week. How should managers be expected to react to that law? What effect should such a law have on the quantity of labor demanded? Why?

2 Does union support of laws outlawing child labor square with the private interests of union members? Should society protect some of its members from some kinds of employment regardless of monetary considerations? Why?

[62] Milton Friedman, "Migrant Workers," *Newsweek* (July 27, 1970): 60

3 How could the minimum wage rate and migrant housing standards be expected to affect the prices of consumer goods? Explain, using supply and demand graphs.

4 Suppose government requires employers to pay a minimum wage of $10 per hour to workers over twenty-two years of age. What effect should such a law have on the employment opportunities and wage rates of persons under twenty-two?

5 Average real wages have increased steadily over the last 100 years. What do you think is the main cause of the increase?

6 Suppose there were a cap put on executive pay by the government. Suppose that "excessive wages" of executives were "excessively" taxed. What would be the effects on wages of workers down the corporate ladder?

CHAPTER 16

Public Choice: Politics in Government And the Workplace

I have no fear, but that the result of our experiment will be, that men may be trusted to govern themselves without a master. Could the contrary be proved, I should conclude, either that there is no God, or that he is a malevolent being.

Thomas Jefferson

Previous chapters have discussed the effects of various government policies on the market system in general and the firm in particular. We looked at government efforts to control the external costs of pollution. We considered the economic impact of price controls and consumer protection laws, for example, on the market for final goods and services. Throughout the analysis we have focused on assessing the economic efficiency of government policy. We said little about how government policy is determined or why government prefers one policy to another.

In this chapter, we will shift our focus to the functioning of government itself. Using economic principles, we will examine the process through which government decisions are made and carried out in a two-party democratic system, and consider its consequences. Today, when government production accounts for a substantial portion of the nation's goods and services, no student of economics can afford to ignore these issues.

A study of the political process is especially important for many MBA students, mainly because a non-trivial amount of your time will be involved with seeking to change one governmental policy or another. Moreover, politics is also endemic to many businesses. Our discussion of the "economics of politics" has various implications for how businesses can be expected to operate, especially those that rely on "participatory management" processes (which are necessarily democratic to one extent or another).

The Central Tendency of a Two-Party System

In a two-party democratic system, elected officials typically take middle-of-the road positions. Winning candidates tend to represent the moderate views of many voters who are neither liberals nor conservatives. For this reason there is generally little difference between Republican and Democratic candidates. Even when the major parties' candidates differ strongly, as Ronald Reagan and Walter Mondale did at the start of their 1984 presidential campaign, they tend to move closer together as the campaign progresses.

Figure 16.1 illustrates politicians' incentives to move toward the center. The bell-shaped curve shows the approximate distribution of voters along the political spectrum. A few voters have views that place them in the wings of the distribution, but most cluster near the center. Assuming that citizens will vote for the candidate who most closely approximates their own political position, a politician who wants to win the election will not choose a position in the wings of the distribution.

Suppose, for instance, that the Republican candidate chooses a position at R_1. The Democratic candidate can easily win the election by taking a position slightly to the left, at D_1. Although the Republican will take all the votes to the right of R_1 and roughly half the votes between R_1 and D_1, the Democrat will take all the votes to the left. Clearly the Democrat will win an overwhelming majority.

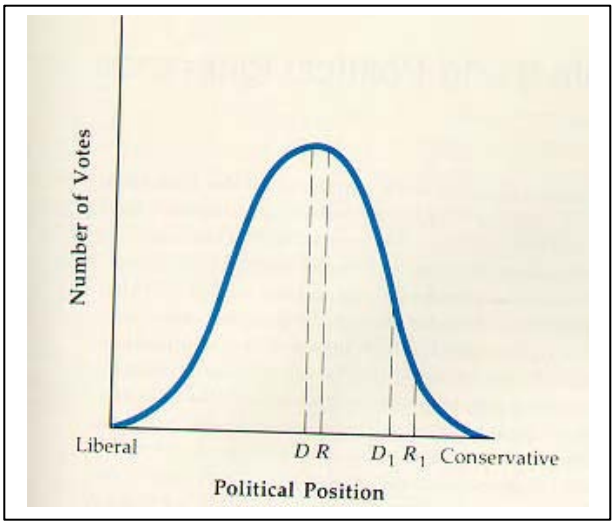


FIGURE 16.1 The Political Spectrum

A political candidate who takes a position in the wings of a voter distribution, such as D_1 or R_1, will win fewer votes than a candidate who moves toward the middle of the distribution. In a two-party election, therefore, both candidates will take middle-of-the-road positions, such as *D* and *R*.

The smart politician, therefore, will choose a position near the middle. Then the opposing candidate must also move to the middle, or accept certain defeat. Suppose, for instance, that the Republican candidate chooses position *R*, but the Democrat remains at D_1. The Republican will take all the votes to the left of *R* and roughly half the votes between *R* and D_1. She will have more than the simple majority needed to beat her Democratic opponent. In short, both candidates will choose political positions in the middle of the distribution.

Politicians can misinterpret the political climate, of course. Even with polls, no one can be certain of the distribution of votes before an election. Just as producers find the optimum production level through trial and error, politicians may suffer several defeats before finding the true center of public opinion. Inevitably, however, political competition will drive them toward the middle of the distribution, where the median voter group resides. The **median voter** is in the middle of the political distribution.

The recent history of presidential elections illustrates how politicians play to the views of the median voter. After an election in which the successful candidate won by a wide margin, the losing party as moved toward the position of the winning party. After Barry Goldwater lost by a wide margin to Lyndon Johnson in 1964, the Republican Party made a deliberate effort to pick a more moderate candidate. As a result, the contest

between Richard Nixon and Hubert Humphrey in 1968 was practically a dead heat. After George McGovern was defeated by Richard Nixon in 1972, Democrats realized they too needed a less extreme candidate. Their choices in 1976 and 1984, Jimmy Carter and Walter Mondale, were more moderate.

In more recent times, after Ronald Reagan soundly defeated Jimmy Carter and Walter Mondale and George Bush beat Michael Dukakis in 1988, the Democrats began what appeared to be a move back toward the center, picking Bill Clinton, a centrist candidate whose policies, in many ways, have been more conservative than were George Bush's.

The Economics of the Voting Rule

So far we have been assuming that a winning candidate must receive more than 50 percent of the vote. Although most issues that confront civic bodies are determined by simple-majority rule, not all collective decisions are made on that basis, nor should they be. Some decisions are too trivial for group consideration. The cost of a bad decision is so small that it is uneconomical to put the question up for debate. Other decisions are too important to be decided by a simple majority. Richard Nixon was elected president with only 43 percent of the popular vote in 1968 (when a third-party candidate, George Wallace, took almost 14 percent), but Nixon's impeachment would have required more than a majority of the Senate and the House of Representatives. In murder cases, juries are required to reach unanimous agreement. In such instances, the cost of a misguided decision is high enough to justify the extra time and trouble required to achieve more than a simple majority.

The voting rule that government follows helps determine the size and scope of government activities. If only a few people need to agree on budgetary proposals, for example, the effect can be to foster big government. Under such an arrangement, small groups can easily pass their proposals, expanding the scope of government activity each time they do so. However, under a voting rule that requires unanimous agreement among voters—a **unanimity rule**—very few proposals will be agreed to or implemented by government. There are very few issues on which everyone can agree, particularly when many people are involved.

A unanimity rule can be exploited by small groups of voters. If everyone's vote is critically important, as it is with a unanimous voting rule, then everyone is in a strategic bargaining position. Anyone can threaten to veto the proposed legislation unless he is given special treatment. Such tactics increase the cost of decision-making.

Government represents the people's collective interest, but the type of voting rule used determines the particular interests it represents and the extent to which it represents them.

The Inefficiencies of Democracy

As a form of government, democracy has some important advantages. It disperses the power of decision making among a large number of people, reducing the influence of individual whim and personal interest. Thus it provides some protection for individual liberties. Democracy also gives political candidates an incentive to seek out and represent voters' interests. Competition for votes forces candidates to reveal what they are willing to do for various interest groups. Like the market system, however, the democratic system has some drawbacks as well. In particular, democracy is less than efficient as a producer of some goods and services.

The fact that the democratic form of government is inefficient in some respects does not mean that we should replace it with another decision-making process, any more than we should replace the market system, which is also plagued by inefficiencies. Instead, we must measure the costs of one type of production against the other, and choose the more efficient means of production in each particular case. We must weigh the cost of externalities in the private market against the cost of inefficiencies in the public sector. Neither system is perfect, so we must choose carefully between them.

Median Voter Preferences

When you buy a good like ice cream in the marketplace, you can decide how much you want. You can adjust the quantity you consume to your individual preferences and your ability to pay. If you join with your neighbors to purchase some public service, however, you must accept whatever quantity of service the collective decision-making process yields. How much of a public good government buys depends not only on citizens' preferences, but also on the voting rule that is used.

Consider police protection, for instance. Perhaps you would prefer to pay higher taxes in return for a larger police force and lower crime rate. Your neighbors might prefer a lower tax rate, a smaller police force, and a higher crime rate, but public goods must be purchased collectively, no matter how the government is organized. If preferences differ, you cannot each have your own way. Under a democracy, the preferences of the median voter group will tend to determine the types and quantities of public goods produced. If you are not a member of that group, the compromise that is necessary to a democracy inflicts a cost on you. You probably will not receive the amount of police protection you want.

The Simple-Majority Voting Rule

Any decision that is made less than unanimously can benefit some people at the expense of others. Because government expenses are shared by all taxpayers, the majority that votes for a project imposes an external cost on the minority that votes against it. Consider a democratic community composed of only five people, each of whom would benefit to some degree from a proposed public park. If the cost of the park, $500, is divided evenly among the five, each will pay a tax of $100. The costs and benefits to

each taxpayer are shown in Table 16.1. Because the total benefits of the project ($550) exceed its total cost ($500), the measure will pass by a vote of three to two, but the majority of three imposes net costs of $50 and $75 on taxpayers D and E.

Table 16.1 Costs and Benefits of a Public Park for Five People

Individuals (1)	Dollar Value of Benefits to Each Person (2)	Tax Levied on Each Person (3)	Net Benefit (+) or Net Cost (-) [(2) - (3)] (4)	Vote For or Against (5)
A	$200	$100	+ $100	For
B	150	100	+ 50	For
C	125	100	+ 25	For
D	50	100	- 50	Against
E	25	100	- 75	Against
Total	$550	$500		

When total benefits exceed total costs, as in this example, decision by majority rule is fairly easy to live with, but sometimes a project passes even though its cost exceeds its benefits. Table 16.2 illustrates such a situation. Again, the $500 cost of a proposed park is shared equally by five people. Total benefits are only $430, but again they are unevenly distributed. Taxpayers A, B, and C each receive benefits that outweigh a $100 tax cost. Thus A, B, and C will pass the project, even though it cannot be justified on economic grounds.

Table 16.2 Costs and Benefits of a Public Park for Five People
Alternative Schedule

Individuals (1)	Dollar Value of Benefits to Each Person (2)	Tax Levied on Each Person (3)	Net Benefit (+) or Net Cost (-) [(2) - (3)] (4)	Vote For or Against (5)
A	$140	$100	+$ 40	For
B	130	100	+ 30	For
C	110	100	+ 10	For
D	50	100	- 50	Against
E	0	100	- 100	Against
Total	$430	$500		

It is conceivable that many different measures, each of whose costs exceed its benefits, could be passed by separate votes under such a system. If all the measures were considered together, however, the package could be defeated. Consider the costs and benefits of three proposed projects—a park, a road, and a school—shown in Table 16.3. If the park is put to a vote by itself, it will receive the majority support from A, B and C. Similarly, the road will pass with the support of A, C, and E, and the school will pass

with the support of C, D, and E. If all three projects are considered together, however, they will be defeated. Voters A, B, and D will reject the package (see column 4).

Table 16.3 Costs and Benefits of a Park, a Road, and a School

	Park (1)			*Road (2)*			*School (3)*			*Total, 3 Projects (4)*		
Individuals	Benefit	Cost	Vote	Benefit	Cost	Vote	Benefit	Cost	Vote	Benefit	Cost	Vote
A	$120	$100	For	$250	$200	For	$50	$400	Against	$420	$700	Against
B	120	100	For	50	200	Against	50	400	Against	220	700	Against
C	120	100	For	250	200	For	500	400	For	870	700	For
D	50	100	Against	50	200	Against	500	400	For	600	700	Against
E	50	100	Against	250	200	For	500	400	For	800	700	For
Total	$460	$500		$750	$1,000		$1,600	$2,000		$2,910	$3,500	

Many if not most measures that come up for a vote in a democratic government benefit society more than they burden it. Moreover, voters in the minority camp can use "logrolling" (vote trading) to defeat some projects that might otherwise pass. For instance, voter A can agree to vote against the park if voter D will vote against the school. Our purpose is simply to demonstrate that, in some instances, the democratic process can be less than cost efficient.

Political Ignorance

In some ways, the lack of an informed citizenry is the most severe problem in a democratic system. The typical voter is not well informed about political issues and candidates. In fact, the average individual's welfare is not perceptibly improved by knowledge of public issues.

A simple experiment will illustrate this point. Ask everyone in your class to write down the name of his or her congressional representative. Then ask them for the name of the opposing candidate in the last election. You may be surprised by the results. In one survey, college juniors and seniors, most of whom had taken several courses in economics, political science, and sociology, were asked how their U.S. senators had voted on some major bills. The students score no better than they would have done by guessing.[1] In the United States, most voters do not even know which party controls Congress,[2] and public opinion polls indicate that most voters greatly underestimate the cost of programs like Social Security.[3]

[1] Richard B. McKenzie, "Political Ignorance: An Empirical Assessment of Educational Remedies," *Frontiers of Economics* (Blacksburg, VA.: University Publications, 1977)

[2] Donald E. Stokes and Warren E. Miller, "Party Government and the Saliency of Congress." *Public Opinion Quarterly* 26 (Winter 1962): 531-546.

[3] Edgar Browning, "Why the Social Insurance Budget Is Too Large in a Democracy," *Economic Inquiry* 13 (September 1974): 373-388

If voters were better informed on legislative proposals and their implications, government might make better decisions. In that sense, political information is a public good that benefits everyone. Nevertheless, as we have seen before, in large groups people have little incentive to contribute anything toward the production of a public good. Their individual contributions simply have little effect on the outcome.

To remain politically free, people must exercise their right to determine who will represent them. The result is that they often cast their votes on the basis of impressions received from newspaper headlines or television commercials—impressions carefully created by advertisers and press secretaries.

Special Interests

The problem of political ignorance is especially acute when the benefits of government programs are spread more or less evenly, so that the benefits to each person are relatively small. Benefits are not always spread evenly: subgroups of voters—farmers, labor unions, or civil servants—often receive more than their proportional share. Members of such groups thus have a special incentive to acquire information on legislative proposals. Farmers can be expected to know more about farm programs than the average voter. Civil servants will keep abreast of proposed pay increases and fringe benefits for government workers, and defense contractors will take a private interest in the military budget.

Congressional representatives, knowing they are being watched by special-interest groups, will tend to cater to their wishes. As a result, government programs will be designed to serve the interest of groups with political clout, not the public as a whole.

Cyclical Majorities

In their personal lives, most people tend to act consistently on the basis of rational goals. If an individual prefers good A to good B, and good B to good C, the rational individual will choose A over C repeatedly. Collective decisions made by majority rule are not always consistent. Consider a community of three people, whose preferences for goods A, B, and C are as follows:

Individual	**Order of Preference**
I	A, B, C
II	B, C, A
III	C, A, B

Supposed these three voters are presented with a choice between successive pairs of goods, A, B, and C. If the choice is between good A and good B, which will be preferred collectively? The answer is A, because individuals I and III both prefer it to B. If A is pitted against C, which will be preferred? The answer is C, because individuals II and III both prefer it to A. Since the group prefers A to B and C to A, one might think it

would prefer C to B, but note that if C and B are put up to a vote, B will win. A cyclical, or revolving majority has developed in this group situation. This phenomenon can lead to continual changes in policy in a government based on collective decision-making.

Although there is no stable majority, the individuals involved are not acting irrationally. People with perfectly consistent personal preferences can make inconsistent collective choices when acting as a group. Fortunately, the larger the number of voters and issues at stake, the less likely a cyclical majority is to develop. Still, citizens of a democratic state should recognize that the political process may generate a series of inconsistent or even contradictory policies.

The Efficiencies of Competition Among Governments

In the private sector, competition among producers keeps prices down and productivity up. A producer who is just one of many knows that any independent attempt to raise prices or lower quality will fail. Customers will switch to other products or buy from other producers, and sales will fall sharply. To avoid being undersold, therefore, the individual producer must minimize its production costs. Only a producer who has no competition—that is a monopolist—can afford to raise the price of a product without fear of losing profits.

These points apply to the public as well as the private sector. The framers of the Constitution, in fact, bore them in mind when they set up the federal government. Recognizing the benefits of competition, they established a system of competing state governments loosely joined in federation. As James Madison Described in *The Federalist* papers, "In a single republic, all the power surrendered by the people is submitted to the administration of a single government: and the usurpations are guarded against by a division of the government into distinct and separate departments."[4]

Under the federal system, the power of local governments is checked not just by citizens' ability to vote, but also by their ability to move somewhere else. If a city government raises its taxes or lowers the quality of its services, residents can go elsewhere, taking with them part of the city's tax base. Of course, many people are reluctant to move, and so government has a measure of monopoly power, but competition among governments affords at least some protection against the abuses of power.

Local competition in government has its drawbacks. Just as in private industry, large governments realize economies of scale in the production of services. Garbage, road, and sewage service can be provided at lower cost on a larger scale. For this reason, it is frequently argued that local governments, especially in metropolitan areas, should consolidate. Moreover, many of the benefits offered by local governments spill over into surrounding areas. For example, people who live just outside San Francisco may benefit from its services, without helping pay for them. One large metropolitan government,

---

[4] Alexander Hamilton, John Jay, and James Madison, *The Federalist: A Commentary on the Constitution of the United States*, no. 51 (New York: Random House, Modern Library edition, 1964), pp. 338-339.

including both city and suburbs, could spread the tax burden over all those who benefit from city services.

Consolidation can be a mixed blessing, however, if it reduces competition among governments. A large government restricts the number and variety of alternatives open to citizens and increases the cost of moving to another locale by increasing the geographical size of its jurisdiction. Consolidation, in other words, can increase government's monopoly power. As long as politicians and government employees pursue only the public interest, no harm may be done. In fact, the people who run government have interests of their own. So the potential for achieving greater efficiency through consolidation could easily be lost in bureaucratic red tape. Studies of consolidation in government are inconclusive, but it seems clear that consolidation proposals should be examined carefully.

The Economics of Government Bureaucracy

Bureaucracy is not limited to government. Large corporations like General Motors and AT&T employ more people than the governments of some nations. They are bigger than the major departments of the federal government—although no company, of course, is as large as the federal government as a whole. Yet corporate bureaucracy tends to work more efficiently than government bureaucracy. The reason may be found in the fact that it pursues one simple objective—profit—that can be easily measured in dollars and cents.

Certainly the reason cannot be that stockholders are better informed than voters. Most stockholders are rationally ignorant or their companies' doings, for the cost of becoming informed outweighs the benefits. Even in very large corporations, however, some individuals hold enough stock to make the acquisition of information a rational act. Often such stockholders sit on the company's board of directors, where their interest in increasing the value of their own shares makes them good representatives of the rest of the stockholders. The crucial point is that this informed stockholder has one relatively simple objective—profit—and can find out relatively easily whether the corporation is meeting it. The voter, on the other hand has a complicated set of objectives and must do considerable digging to find out whether they are being met.

Because most corporations function in competitive markets, the stockholder's drive toward profit is reinforced. General Motors knows that its customers may switch to Toyota if it offers them a better deal. In fact, stockholders can sell their General Motors stock and buy stock in Toyota. Thus corporate executives make decisions on the basis of the consumer's well being—not because they wish to serve the public good but because they want to make money.

Government bureaucracies, on the other hand, tend to produce public goods and services for which there is no competition. No built-in efficiencies guard the taxpayer's interests in a government bureaucracy. Both government bureaucrats and corporate executives base their decisions on their own interests, not those of society, but competition ensures that the interests of corporate decision makers coincide with those of consumers. No such safeguards govern the operations of government bureaucracies. Bureaucracies are constrained by political, as opposed to market, forces.

From the economist's point of view, one of the advantages of the profit-maximizing goal of competitive business is that it enables predictions. Although some business people pursue other goals—personal income, power, respect in the business—their behavior can generally be well explained in terms of the single objective, profit. There is no single goal like profit that drives the government bureaucracy. Different bureaucracies pursue different objectives. We do not have time or space to consider all the possible objectives of bureaucracy, but we will touch on three: monopolistic profit maximization, size maximization, and waste maximization.

Profit Maximization

Assume that police protection can be produced at a constant marginal cost, as shown by the horizontal marginal cost curve in Figure 16.2. The demand for police protection is shown by the downward-sloping demand curve *D*. If individuals could purchase police service competitively at a constant price of P_1, the optimum amount of police service would be Q_2, the amount at which the marginal cost of the last unit of police service equals its marginal benefit. The total cost would be P_1 x Q_2 (or the area $0P_1a\ Q_2$), leaving a consumer surplus equal to the triangular area P_1P_3a.

Police protection is usually delivered by regional monopolies, however. That is, all police services in an area are supplied by one organization. These regional monopolies have their own goals and their own decision-making process, which do not necessarily match the individual taxpayers'. If police service must be purchased from such a profit-maximizing monopoly, service will be produced to the point where the marginal cost of the last unit produced equals its marginal revenue: Q_1. The monopolist will set that quantity above cost at price P_2, making a profit equal to the rectangular area $P_1\ P_2ed$.

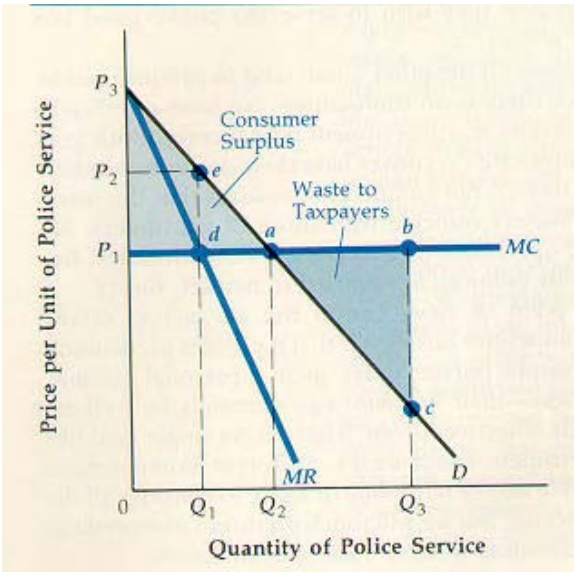


FIGURE 16.2 Bureaucratic Profit Maximization

Given the demand for police service, *D*, and the marginal cost of providing it, *MC*, the optimum quantity of police service is Q_2. A monopolistic police department interested maximizing its profits will supply only Q_1 service at a price of P_2, however. (A monopolistic bureaucracy interested in maximizing its size would expand police service to Q_3.)

At the monopolized production level, there is still some surplus—the triangular area $P_2\ P_3e$—left for consumers, but they are worse off than under competitive market conditions. They get less police protection (Q_1 instead of Q_2) for a higher price (P_2 instead of P_1).

This analysis presumes that the police are capable of concealing their costs. If taxpayers know that P_2 is an unnecessarily high price, the outcome will be the same as under competition. They will force the police to produce Q_2 protection for a price of P_1.

Size Maximization

In fact, a government bureaucracy is unlikely to take profit as its overriding objective, if only because bureaucrats do not get to pocket the profit. Instead, government monopolies may try to maximize the size of their operations. For if a bureaucracy expands, those who work for it will have more chance of promotion. Their power, influence, and public standing will improve, along with their offices and equipment.

What level of protection will a police department produce under such conditions? Instead of providing Q_1 service and misrepresenting its cost at P_2, it will probably provide Q_3 service—more than taxpayers desire—at the true price of P_1. The bill will be P_1 x Q_3, or the area $0P_1b\ Q_3$ in Figure 16.3. Note that the net waste to taxpayers, shown by the shaded area *abc*, exactly equals the consumer surplus, P_1P_3a. By extending service to Q_3, the police have squeezed out the entire consumer surplus and spent it on themselves.

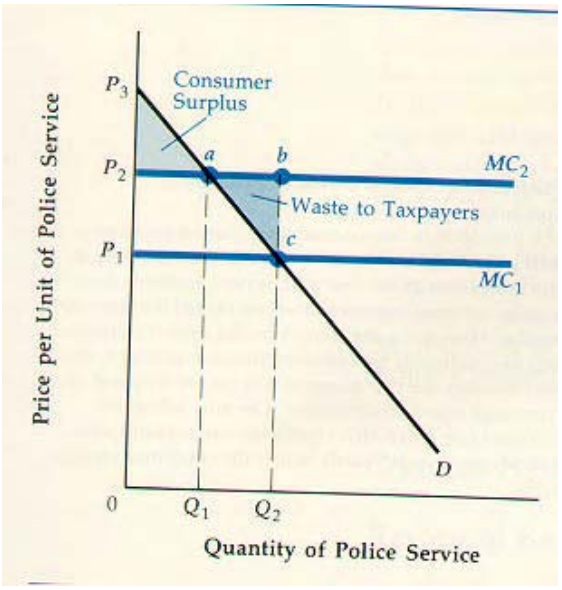


FIGURE 16.3 Bureaucratic Waste Maximization

Given a demand for police service *D* and a marginal cost of providing it MC_1, the optimum quantity of police service will be Q_2. A monopolistic bureaucracy, however, may seek to maximize waste by inflating its costs to MC_2. It will supply Q_2 units of police protection at a tax price of P_2 instead of P_1. The shaded area *abc* shows the waste created, which exactly equals the consumer surplus P_2P_3a.

Waste Maximization

Instead of maximizing the amount of service they offer, bureaucrats may choose to maximize waste. They can increase their salaries, improve their working conditions, or

reduce their workloads. All such changes increase the cost of providing a given amount of service.

Figure 16.3 shows how far a bureau can go in increasing the cost of, or budget for, its services. The marginal cost curve MC_1 is the minimum cost of providing additional police protection. The optimum quantity of police protection is therefore Q_2, the same as in Figure 16.2, but if the police pad their costs, the marginal cost curve will shift up to MC_2. The bureau's budget climbs from P_1 x Q_2 to P_2 x Q_2. Note that beyond Q_1, the marginal cost of additional police service is now greater than its marginal benefit, indicated by the demand curve. Again, the police are wasting taxpayers' money, as shown by the shaded triangular area *abc*. By moving their cost curve to MC_2, they have managed to extract all the consumer surplus (shown by the triangular area P_2P_3a) and to spend it on unnecessary frills.

In real life, most bureaucratic monopolies may pursue both size maximization and waste maximization. For each unit of service they provide, they will try to expand both the size of their operation and the funds spent on it—but they do have to make tradeoffs between the two objectives. Whenever they expand their size, they must forgo a certain amount of expansion in their cost per unit of service. There is, after all, only so much consumer surplus that can be extracted from the system.

Figure 16.4 shows one possible combination of size and budget maximization. In this case the department chooses to expand its service from Q_1 to Q_2. Having done so, it can expand its cost per unit only to MC_2. Again, the shaded triangular area that indicates waste, *abc*, just equals the consumer surplus P_2P_3a.

Fortunately government bureaucracies do not usually achieve perfect maximization of size or waste. For one thing, most legislatures have at least some information about the production costs of various services, and bureaucrats may not be willing to do the hard work necessary to exploit their position fully. If bureaucracy does not manage to capture the entire consumer surplus, citizens will realize some net benefit from their investment.

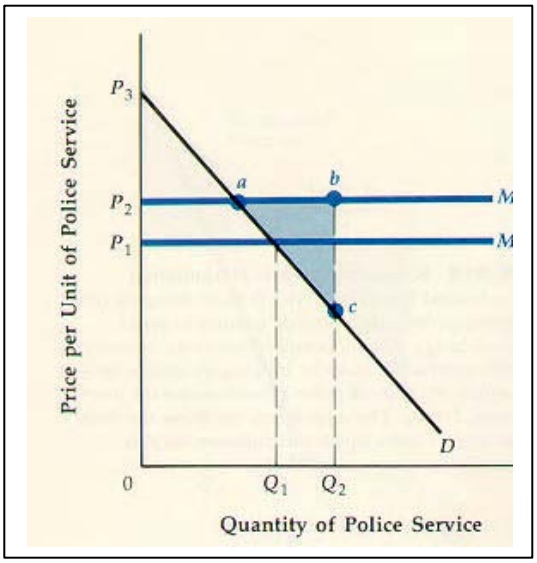


FIGURE 16.4 Size and Waste Maximization Combined

The monopolistic bureaucracy may choose to increase both its size and the cost of its service. Any increase to one must come at the cost of the other, however, for together the two increases must not exceed the consumer surplus. Here net waste, shown by the shaded triangular area *abc*, is divided between size and cost increases. The area between the two marginal cost curves MC_1 and MC_2 represents waste maximization. The are below the marginal cost curve MC_1 represents size maximization. The whole area *abc* exactly equals the consumer surplus P_2P_3a.

Making Bureaucracy More Competitive

What can be done to make government bureaucracy more efficient? Perhaps the development of managerial expertise at the congressional level would encourage more accurate measurement of the costs and benefits of government programs. Cost-benefit analysis alone, however, will not necessarily help. As long as special-interest groups, including those of government employees, exist, the potential for waste can be substantial.

A better solution to bureaucratic inefficiency may be to increase competition in the public sector. In the private marketplace, buyers do not attempt to discover the production costs of the companies they buy from. They simply compare the various products offered, in terms of price and quality, and choose the best value for their money. A monopoly of any kind, of course, makes that task difficult if not impossible, but the existence of even one competitor for a government bureaucracy's services would allow some comparison of costs. The more different sources of a service, the flatter the demand curve faced by each source, and the more efficient it must be to stay in business.

How exactly can competition be introduced into bureaucracy? First, proposals to consolidate departments should be carefully scrutinized. What appears to be wasteful duplication may actually be a source of competition in the provision of service. In the private sector, we would not expect the consolidation of General Motors, Ford and Chrysler to improve the efficiency of the auto industry. If anything, we would favor the breakup of the large firms into separate, competing companies. Why then should we merge the sanitation departments of three separate cities?

A second way to increase the competitiveness of government services is to contract for them with private producers. Many government activities that must be publicly financed need not necessarily be publicly produced. In the United States, highways are usually built by private companies but repaired and maintained by government. Competitive provision of maintenance as well as construction might reduce costs. Other services that might be "privatized" are fire protection, garbage collection, and education.

Finally, competition can be increased simply by dividing a bureaucracy into several smaller departments with separate budgets, thus increasing competition. Such a change would reduce the costs citizens must bear to move to an area that offers better or cheaper government services. The loss (or threat of loss) of constituents can put pressure on government to improve its performance.

MANAGER'S CORNER: **Why Professors Have Tenure and Business People Don't**

Tenure is nothing short of a Holy Grail for newly employed assistant professors in the country's colleges and universities. Without tenure, faculty members must, as a general rule, be dismissed after seven years of service, which means they must seek other academic employment or retreat from academic life. With tenure, professors have the

equivalent of lifetime employment. Rarely are they fired by their academies, even if they become incompetent at teaching and/or researching.

Business people rarely, if ever, have the type of tenure protection that professors do. Why the different treatment? Is it that universities are stupid, bureaucratic organizations in which professors are able to obtain special treatment? Maybe so, but we would like to think not. (Indeed, we think our universities have shown great wisdom in granting us both tenure in our current positions, from which we could not be dislodged with anything short of a direct nuclear hit!) We suggest that our explanation for why professors have tenure will help us understand why some form of tenure will gradually find its way into businesses that have begun to rely progressively more on "participatory management" (with low-ranking managers and line workers having a greater say in how the business is conducted).

The Nature of Tenure

Professors do not, of course, have complete protection from dismissal, and the potential for being fired is surely greater than that reflected in the number of actual firings. However, when professors are fired it is generally for causes unrelated to their professional competence. The most likely reasons for dismissal are "moral turpitude" (which is academic code for sexual indiscretions with students) and financial exigencies (in which case, typically, whole departments are eliminated).

Most proponents and opponents of academic tenure like to think of it in emotional terms: "Tenure is stupid" or "Tenure ensures our constitutional rights." We would like to suggest that tenure be treated as a part of the employment relationship. It amounts to an employment contract provision that specifies, in effect, that the holder cannot easily be fired. To that extent, tenure provides some employment security, but by no means perfect security. A university may not be able to fire a faculty member quickly, but it can repeatedly deny salary increases and gradually increase teaching loads until the faculty member "chooses" to leave.[5]

Clearly, tenure has costs that must be suffered by the various constituencies of universities. Professors sometimes do exploit tenure by shirking their duties in the classroom, in their research, and in their service to their universities. However, tenure is not the only contract provision that has costs. Health insurance (as well as a host of other fringe benefits) for professors imposes costs directly on colleges or universities and indirectly on students. Nonetheless, health insurance costs continue to be covered by universities because the benefits matter too, not just the costs. Health insurance survives as a fringe benefit because it represents, on balance, a mutually beneficial trade for the various constituencies of universities. Universities (which can buy group insurance policies more cheaply than individual faculty members) are able to lower their wage bills by more than enough to cover the insurance costs because they provide health insurance.

---

[5] Accordingly, the degree of protection tenure affords is a function of such variables as the inflation rate. That is, the higher the inflation rate, the more quickly the real value of the professor's salary will erode each time a raise is denied.

By the same token, professors pay for tenure just as they do other fringe benefits; presumably tenure is worth more to them than the value of the foregone wages.

Why tenure? Any reasonable answer must start with the recognition that academic labor markets are tolerably, if not highly, competitive, with thousands of employers and hundreds of thousands of professors, and wages and fringe benefits respond fairly well to market conditions. If, in fact, tenure were not a mutually beneficial trade between employers and employees, universities -- which are constantly in search of more highly qualified students, faculty at lower costs, and higher recognition for their programs -- would be expected to alter the employment contract, modify the tenure provision, increase other forms of payment, and lower overall university costs.[6]

The analysis continues with the recognition that jobs vary in difficulty, in time and skills required, and in satisfaction. "Bosses" can define many jobs, and they are generally quite capable of evaluating the performance of those they hire for these jobs. In response to sales, for example, supervisors in fast food restaurants can determine not only how many hamburgers to cook but also how many employees are needed to flip those hamburgers (and assemble the different types of hamburgers). Where work is relatively simple and routine, we would expect it to be defined by and evaluated within an authoritarian/hierarchical governance structure of firms, as is generally true in the fast-food industry.

Academic work is substantially different, partially because many forms of the work are highly sophisticated, its pursuit cannot be observed directly and easily (given the reliance on thinking skills), and it involves a search for new knowledge which, when found, is transmitted to professional and student audiences. (Academic work is not the only form of work that is heavily weighted with these attributes, a point that will be reconsidered later.) Academic supervisors may know in broad terms what a "degree" should be and how "majors" should be constituted at any given time. However, they must rely ultimately and extensively (but not necessarily completely) on their workers/professors to define their own specific research and classroom curriculums and to change the content of degrees and majors as knowledge in each field evolves. Academic administrators employ people to conduct research and explore uncharted avenues of knowledge that the administrators themselves cannot conduct or explore because they lack knowledge of a field, have no time, or are not so inclined to do so.

Fast-food restaurants can be governed extensively (but not exclusively) by commands from supervisors, and there is an obvious reason why this is possible. Again, the goods and services produced are easily valued and sold, with little delay between the time they are produced and the time the value is realized and easily evaluated. Workers in such market environments would be inclined to see supervisors as people who

---

[6] Granted, tenure may be required by accrediting associations. However, there is no reason that groups of universities could not operate outside accrediting associations or organize their own accrediting associations without the tenure provision -- if tenure were, on balance, a significant impairment to academic goals. In many respects, the accrediting association rules can be defended on the same competitive grounds that recruiting rules of the National Collegiate Athletic Association are defended. See Richard B. McKenzie and T. Sullivan, "The NCAA as a Cartel: An Economic and Legal Reinterpretation," Antitrust Bulletin, no. 3 (1987), 373-399.

increase the income of stockholders *and* workers mainly by reducing the extent to which workers shirk their agreed upon duties.

Academe, however, is a type of business that tends to be worker managed and controlled, at least in many significant ways. This aspect of the academic marketplace solves many decision-making problems but introduces other serious problems of unstable, if not volatile and uncertain, decisions over time and circumstances from which professors will seek contractual protection. Professors are extensively called upon to determine what their firms (universities) produce (what research will be done, what courses are required, and what will be the contents of the various courses, even who will be taught). In addition, they help to determine who is hired to teach identified courses and undertake related research, how workers are evaluated, and when they are fired.

Our argument can be stated without using the examples of fast food and academe, but those examples enable us to deduce a managerial principle of sorts: the simpler it is to accomplish a job, the more likely it is that managerial control will be delegated to a supervisor. The more sophisticated, esoteric, and varied the job to be done, the more likely managerial control will be relegated to the workers themselves and the more democratic the decision-making will be.[7]

Again, why academic tenure? We think the forces of supply and demand for tenure are at work. Economists have argued that universities have reason to "supply" tenure.[8] The reason given: professors are called upon to select new members, which stands in sharp contrast to the way similar decisions are made in business as well as in sports. In baseball, the owners through their agents determine who plays what position on the team. Baseball is, in this sense, "owner managed." In academe, the incumbent professors select the team members and determine which positions they play. Academe is, in this sense, "labor managed."

In baseball, the owners' positions are improved when they select "better players." On the other hand, in academe, without tenure, the position of the incumbent decision-makers could be undermined by their selection of "better professors," those who could teach better and undertake more and higher quality research for publication in higher-ranking journals.[9] Weaker department members would fear that their future livelihoods (as well as prestige) would be undermined by revelation of their honest evaluations of candidates who are better than themselves.

Thus, tenure can be construed as a means employed by university administrators and board members -- who must delegate decision-making authority to the faculty but

---

[7] Of course, not all academic environments share the same goals or face the same constraints. Some universities view pushing back the frontiers of knowledge as central to their mission, while others are intent on transmitting the received and accepted wisdom of the times, if not the ages. Some universities are concerned mainly with promoting the pursuit of usable (private goods) knowledge, that which has a reasonable probability of being turned into salable products, while other universities are interested in promoting research the benefits of which are truly public, if any value at all can be ascertained.

[8] H. L. Carmichael, "Incentives in Academics: Why Is There Tenure?" Journal of Political Economy, vol. 96, no. 2 (1988), pp. 453-472.

[9] "Loosely, tenure is necessary," Carmichael concludes, "because without it incumbents would never be willing to hire people who turn out to be better than themselves" (Ibid., 1988, p. 454).

who still want to elevate the quality of what is done at their universities -- to induce faculty members to honestly judge the potential of the new recruits. In effect, university officials and board members strike a bargain (with varying degrees of credibility) with their professor-decision-makers: If you select new recruits who are better than you are, you will not be fired.

Universities have reason to *supply* tenure, but what reason do professors have to *demand* it? We don't buy the argument that most faculty members want to be protected from the broader political forces outside the ivy-covered walls of their universities. Too few faculty members ever go public with their work or say anything controversial in their classes for them to want to give up very much for such protection from external forces. Rather, we believe tenure is designed to protect professors from their colleagues, acting alone or in political coalitions, in a labor-managed work environment operating under the rules of academic democracy. That is, faculty members demand tenure so that there will be little or no incentive for other faculty members to run them out of the decision-making unit.

Academic work is often full of strife, and the reasons are embedded in the nature of the work and the way work is evaluated and rewarded, a point one of the authors has discussed in detail elsewhere.[10] Suffice it to say here that tenure is a means of putting some minimum limits on political infighting. It increases the costs predatory faculty members must incur to be successful in having more productive colleagues dismissed. More importantly, academic decisions on the worth of colleagues and their work are often made by the rules of consensus or democracy among existing incumbents.

Certainly, most professors understand both the esoteric nature of their work and the problems of short-term evaluations. At the same time, they understand that in an academic democracy, ever-changing groups of colleagues have a say in how the work of each professor is evaluated. They recognize implicitly, if not explicitly, that how their work is evaluated by a changing group of colleagues can depend, at the time, on what their work is being compared with. A microeconomics scholar can appreciate the fact that the relative ranking of his or her research depends upon whether it is being judged relative to the work of macro or public policy scholars.

In addition, professors understand that the relative standing of their positions and ranking of their research can change over time with changes in the cast of decision-makers, who are likely to adjust their assessments from time to time. The ranking of their research can also change with shifts in the relative merit department members assign different types and forms of academic work. For example, a macro person understands that even though his or her publications may now be highly valued (relatively) within the department, the ranking can easily change, because changes occur in the way evaluations are made, existing department members periodically reassess the relative worth of different types of work, and the cast of decision-makers changes. When the decision-making unit is multi-disciplinary, shifts in the relative assessments of the worth of individual professors' work in the different disciplines can fluctuate even more

[10] Richard B. McKenzie, "The Economic Basis of Departmental Discord in Academe," Social Science Quarterly, no. 1 (1979), pp. 653-664.

dramatically, given that each professor is likely to have allegiance first to his or her own discipline and then to other closely related disciplines.

Within schools of business, for example, accounting faculty members may have, on the margin, an incentive to depreciate the work of marketing professors, given that such depreciation may shift positions to accounting -- and vice versa. Even more fundamentally, organizational theorists steeped in behavioral psychology may have an incentive to depreciate the work of professors in finance -- which is grounded in economics -- given that negative shifts in the relative evaluation of economic-based work can marginally improve the chances of positions being shifted to, say, accounting. Like-minded faculty members can be expected to coalesce to increase their political effectiveness in shaping decisions that can, in turn, inspire the formation of other coalitions, thus motivating all coalitions to increase their efforts. The inherent instability of coalitions can, of course, jeopardize anyone's job security and long-term gains.

Professors have understandable reasons for demanding tenure. One is that the esoteric nature of their work (which they may undertake at the behest of their universities) may diminish the market value of their skills because the narrow focus of their work *might* not translate into alternative future job opportunities in the market place. Another reason is that there are political problems inherent within all democratic processes, and professors want, in effect, to be protected from the process and from their colleagues. If their work is intensely specialized, they want some assurance of job security to protect against the changing assessments by ever-changing majorities. Universities can be seen as willing to provide tenure because they must delegate decision-making power to those who have the requisite knowledge and information of different disciplines if they want faculty members to specialize their efforts. Universities also realize, given the nature of academic democracy and the threat it poses, that faculty members have inherent reasons for demanding tenure, and these make it possible to recoup the cost of tenure by reducing professorial wages to less than what they would have to be if the professors did not share a need for job security.

Of course, this line of analysis leads to a number of deductions:

- If the work of professors were less specialized, professors would be less inclined to demand tenure. For example, in colleges in which the emphasis is on teaching rather than research, tenure would be less prevalent, or less protective.
- As a group of decision-makers or a discipline becomes more stable, we would expect faculty to consider tenure less important and to be less willing to forgo wages and other fringe benefits to obtain tenure.
- If there is a close to even split on democratic decisions related to employment, merit raises, and even tenure, faculty members will assign more value to tenure, given that a more or less evenly split vote may change with slight shifts in the composition of the decision-makers.
- The further below market are the wages of faculty during the probation period and the further above market are wages after tenure, the more valuable tenure is to faculty members.

- As the diversity within a decision-making unit increases (more disciplines included with more divergent views on how analyses should be organized and pursued), the demand for tenure will increase.
- Should universities become more constrained in their capacities to fund established faculty positions, tenure may be perceived as even more valuable. Financial exigencies can translate into the loss of faculty positions (with non-tenured positions becoming prime targets), so it should not be surprising that faculty will seek with greater diligence to redistribute remaining positions and rents. It also means universities will probably have to spend considerable resources seeking to instill academic values -- not the least of which will be the pursuit of honest dealings and academic excellence. This emphasis may cause faculty members to shun an important incentive inherent in the political process (especially in large group settings), that is, the tendency to pursue strictly private objectives at the expense of larger university goals.[11]

Why Business People Don't Have Tenure

If professors have tenure, why don't business people have provision for the same kind of job security? The quick answer to that question is that businesses, unlike universities, typically are not labor managed. (Those that are like universities should be expected to use some form of tenure.) As noted, in business, goals are usually well defined. Perhaps more importantly, success can usually be identified with relative ease by using an agreed-upon measure, that is, profit (or the expected profit stream captured in the market prices of traded securities). The owners, who are residual claimants, have an interest in maintaining the firm's focus on profits. Moreover, people who work for businesses tend to have a stake in honest evaluations of potential employees, given that their decisions on "better" recruits can increase the firm's profits and the incomes and job security of all parties.

Admittedly, real-world businesses do not always adhere to the process as described. They use, to a greater or lesser degree, participatory forms of management, and for some businesses, profit is not always the sole or highest priority goal. "Office politics" is a nontrivial concern in many firms. The point is, however, that in business there is not as great a need for tenure as exists within academe; employees in businesses do not have the incentive to demand tenure that professors have, primarily because these employees do not experience the problems inherent in democratic management that derive from imprecise and shifting goals and from esoteric and ill-defined research projects. Tenure is seldom found in firms, for the simple reason that in business, employers and employees cannot make mutually beneficial trades (similar to those made in tenure arrangements).

Now, let's suppose that political institutions and problems were as well entrenched in a firm as they are in academe, to the point of significantly undercutting

[11] As Miller (1992) has shown, the benefits of "corporate organization" eventually break down when the parties follow completely rational, individualistic precepts [Gary J. Miller, Managerial Dilemmas: The Political Economy of Hierarchy (New York: Cambridge University Press, 1992)].

firm profits. What would happen? Clearly, some smart coalition of managers or outside investors would see a potential for increasing their wealth. They would buy the firm's stock at a low price depressed by the political encumbrances and reform management practices, suppressing the power of destructive politics and refocusing the managers' and workers' attention on the bottom line. They would clarify the extent to which the workers' long-run gains would be a function of their contributions to profits. The price of the stock could then rise. Voila! The takeover investors would have a wealth increase, and the workers would have less need for tenure, as professors know that form of job protection.

Tenure as a Tournament

We also suggest that the granting of tenure can be seen as another form of the *tournament* we have discussed earlier in other contexts. Tenure decisions are a way of allowing faculty members to reveal their skills. An employer cannot depend on a potential employee to be fully objective or honest in presenting his or her qualifications. The graduate school records of new doctorates provide useful information on which to base judgments of potential recruits for success as university teachers and researchers. However, such records are of limited worth in instances where a professor's research is at the frontier of knowledge in his or her discipline. The correlation between a person's performance as a student, as a prospective professor, as a teacher, and as a researcher is, at best, imperfect.

In order to induce promising faculty members to accurately assess their abilities and to confess their limits, the competitors (new assistant professors) are effectively told that only some among them will be promoted and retained. Since standards for tenure differ from one university to another, universities offer prospective faculty members an opportunity to, in effect, self-select and go to a university where they think they are likely to make the tenure grade. The prospects of being denied tenure will cause many (but certainly not all) weak candidates to avoid universities with tough tenure standards, given the probability that they would have to accept wages well below market during the probation period. The lost wages amount to an investment that probably will not be repaid with interest (in terms of wages above the market after the probation period when tenure is acquired). Thus, the tenure tournaments can reduce to some extent the costs universities incur in gathering information and making decisions, because they force recruits to be somewhat more honest in their claims.

Competition for the limited number of "prized positions" often will drive new faculty members to exert a level of effort and produce a level of output that exceeds the value of their current compensation. To induce prospective faculty to exert the amount of effort necessary to be ability revealing, universities must offer a "prize" that potential recruits consider worth the effort. That is, the recruits must expect the future (discounted) reward to compensate them for the extra effort they expend in the tournament and for the risk associated with not "winning." One approach universities can use to encourage recruits to exert a reasonable level of effort in the competition is to offer those who win the prospect of substantially greater compensation in the future (at least enough to repay the costs of assumed risk and of interest lost on delayed

compensation). Another approach that offers future compensation as an incentive is to increase the security of continued employment and compensation once the tournament has ended and the winners have been determined. That is, tenure can be offered as the "prize."

In the absence of tenure (or some similar device), universities would find it difficult to make a credible commitment that prospective recruits, who make the necessary competitive investment during the probationary period by accepting below-market wages for above-market effort, will receive an income stream that compensates them for all costs, including the required risks. We have stressed the instability inherent in academic democracies that, by its nature, reduces the credibility of virtually every commitment universities might want to make at employment time. Tenure is a practical means universities use to provide a reasonable level of job security -- to make a credible commitment -- that is, to overcome institutional instabilities and thereby enable them to pick the "best" professors for continued employment. At the same time, tenure is part of a mutually beneficial trade between new professors and their universities, primarily because it is a feature of the employment contract that new self-selected faculty members will demand before they agree to participate actively and honestly (in the sense that they will reveal the limits of their true abilities) in what amounts to a risky and underpaid employment tournament, albeit short-run.[12]

After all is said and done, tenure is nothing more than another contract provision that faculty members prize, universities provide -- and just about everyone else criticizes. Business people could also have tenure. All they would have to do is "pay" for it in terms of lost wages. However, business people typically don't have the same strong reasons for wanting tenure as do professors. Tenure survives in the academies of the country mainly because faculty members aggressively demand it (even those who believe strongly in the value of markets) and because universities voluntarily negotiate it. Tenure's long-term survival and the competitiveness of university labor markets suggest that the trade is mutually beneficial.

Concluding Comments

This chapter has used cost-benefit analysis to develop an economic model of government. In government as well as private industry, producers in a monopolistic market position will tend to exploit the lack of competition for their service. A government bureau that has no competitors is in an enviable bargaining position vis-à-vis legislators and taxpayers. As the sole producer of a service, it can charge higher prices and deliver poorer service than competitive producers would.

---

[12] After tenure is awarded, faculty efforts should be expected to decline, while, at the same time, their pay rises. In the midst of the tournament, the new faculty members will exert unduly high amounts of effort, simply because of the prospect of being rewarded in the future by higher pay and greater job security. Also, the rise in compensation and fall in effort that accompany tenure may correlate with the fact that the added money makes it possible for faculty members to buy more of most things, including great leisure (or leisure-time activities). If we did not expect new faculty members to anticipate relaxing somewhat after attaining tenure and enjoy, to a degree, being "overpaid," we could not expect the tenure tournament to be effective as a means to an end, which is disclosure of the limits of new faculty members' true abilities.

In many cases, then, the performance of government bureaucracies can be improved by the introduction of competition for their services. Where possible, alternative sources of a government-provided good or service should be encouraged. If government bureaus have to compete with other producers by lowering their prices or increasing the quality of their service, they will be forced, like private producers, to reveal not just what they want to do, but the limit of what they will do for the consumer's business.

The democratic system provides checks and balances to control the exploitation of power in government. Voters can vote not to re-elect officeholders who abuse the public trust. They may not do so reliably, however, because of imperfect information. The fact that democracy is not a completely efficient system does not mean that a non-democratic form of government is preferable. We have noted, however, that people will also seek protections from the problems intrinsic to democratic governance. They can do this with constitutional restrictions on what governments can do. Inside firms, workers can protect themselves from workplace democracies through contract restrictions like tenure. Owners of firms need to be mindful of the fact that if they move toward "participatory management," they will have to provide worker protections from the majorities' abuse of democratic governance in the workplace, or else the firms will have to pay higher wages.

Review Questions

1. Is it desirable, in your opinion, that government generally adopts policies intended to please the median voter group? Why or why not?
2. It is sometimes said that a rational decision must be based on perfect information. Would it be rational for a voter to acquire perfect information about politics? Would it be possible?
3. What effect does increased competition have on the slope of an individual firm's demand curve? Why? How does a change in the slope of a firm's demand curve affect its efficiency? How do these effects apply to government bureaucracy?
4. "Competition forces producers to reveal what they are willing to do at the limit, not just what they want to do." How does this statement apply to government bureaucracy, and to legislators' ability to control it?
5. Write down all the government-provided services you can think of. Which of them *must* be provided by government bureaucracy? Which could be provided through competitive contract? Why?
6. When would workers want and don't want democratic governance in the workplace?

READING: ***The Mathematics of Voting and Political Ignorance***

Gordon Tullock, University of Arizona

Public problems are normally more important than private problems, but the decision by any individual on a private problem is likely to be more important than his decision on a public problem, simply because most people are not so situated that their decision on public matters makes very much difference. It is rational, therefore, for the average family to put a great deal more thought and investigation into a decision such as what car to buy than into a decision on voting for president. As far as we can tell, families, in fact, act quite rationally in this matter, and the average family devotes almost no time to becoming informed on political matters but will carefully consider the alternatives when buying a car. Why is that the case?

In order to address the question we need first to ask a more basic question: What is the payoff to the individual from voting? Assume that you are in possession of some information and have decided that you favor the Democratic Party or, if is a primary, some particular candidate. The payoff could be computed from the following expression:

$$BDA - C_v = \boldsymbol{P}$$

B = benefit expected to be derived from success of your party or candidate

D = likelihood that your vote will make a difference

A = your estimate of the accuracy of your judgement ($-1<A<+1$)

C_v = cost of voting

P = payoff

Certain aspects of this expression deserve a little further discussion. The B refers, of course, not to the absolute advantage of having one party or candidate in office, but the difference between the candidate and his or her opponent. The factor labeled A, the estimate of the accuracy of the voter's judgement, is included here because we are preparing to consider the amount of information held by the individual, and the principal effect of being better informed is that your judgement is more likely to be correct. The factor labeled A can take any value from minus 1, which represents a certainty that the judgements will be wrong, to a plus 1, which indicates that the voter is sure he or she is right. The choice of this rather unusual way of presenting what is really a probability figure is due solely to its use in the particular equation, not to any desire to change the probability notational scheme. For the equation to give the right answer, it is necessary that A have a value of zero when the individual thinks that he has a fifty-fifty chance of being right.

The factor labeled D is the likelihood that an individual's vote will make a difference in the election; that is, the probability that the result if he were to vote would be different than it would be if her were not to vote. For an American presidential election, this is less than one in 10 million. C_v is the cost, in money and convenience, of voting. For some people, of course, it may be negative. They may get pleasure, or at least the negative benefit of relief of social pressure, from voting. If we view voting as an instrumental act, however—something we do not because it gives us pleasure directly but because we expect it to lead to some desirable goal—then our decision to vote or not will depend on weighing the costs and benefits.

Let us put a few figures into our expression. Suppose I feel that the election of the "right" candidate as president is worth $10,000 to me. I think I am apt to be right three times out of four, so the value of A will be .5, D will be figured as .000,000,1. Assuming that my cost of voting is $1.00, the expression gives ($10,000 x .5 x .000,000,1) - $1.00 = $.9995. It follows from this that I should not bother voting.

It will, however, be worthwhile to consider a few variations on the expression. In the first place, it is frequently argued that this line of reasoning would lead to no one voting. This is not true. If people began making these computations and then refraining from voting, this would raise the value of D, since

the fewer the voters, the more likely that any given vote will affect the outcome. As more and more people stopped voting, *D* would continue to rise until the left side of the expression equaled the right. At this equilibrium there would be no reason for nonvoters to begin to vote or for voters to stop. Presumably the people voting would be those among the population who were most interested in politics, since *D* would have the same value for everyone but ($B \times A$) would approximate a positive function of political interest.

The equation, if it is thought to be in any way descriptive of the real world, would imply that people would be more likely to vote in close elections. This hypothesis has been tested and found to be correct.

Let us now complicate our model. An additional factor, C_i, the cost of obtaining information, has been included in the first equation.

$$BDA - C_v - C_I = P$$

This, of course, the cost of obtaining additional information, since the voter will have at least some information on the issues as a result of his contact with the mass media. Of course, *A* is a function of information ($A = f(I)$), and hence each increase in information held will increase *A* and thus raise both the benefits and the costs. The problem for the rational individual contemplating whether or not he or she should vote would be whether there are any values of C_I that would lead to a positive value payoff.

Suppose, for example, that the investment of $100.00 (mostly in the form of leisure forgone) in obtaining more information would raise the value of *A* from .5 to .8. Using the same amounts for the other values as we used previously, *P* = -$100.9992. Clearly, this is even worse than the original outcome. Furthermore, these figures are realistic. The cost of obtaining enough information to significantly improve your vote is apt to very much outweigh the effect of the improvement. This is particularly true for the average voter, who does not have much experience or skill in research and who would put a particularly high negative evaluation on the time spent in this way.

A further implication of our reasoning must be pointed out. There may be social pressures that make it wise for the individual to make the rather small investment necessary for voting. In terms of our equation, C_v may be negative. In these cases, voting would always be rational. Becoming adequately informed, however, is much more expensive. Further, it is not as easy for your neighbors (or your conscience) to see whether you have or have not put enough thought into your choice. Thus, it would almost never be rational to engage in much study in order to cast a "well-informed" vote. For certain people (and presumably most readers of this book will fall within this category) *A* may already be quite high. For intellectuals interested in politics, the amount of information acquired about the different issues for reasons having nothing to do with voting may be quite great. Further, for this group of people, the value put on the well being of others may be higher than in the rest of the population. It may be, then, that these people would get a positive payoff from voting even though the average citizen would get negative returns from taking the same action. Thus, for many of the readers of this book, voting may be rational. I have my doubts, however. The value put on the well being of others must be extremely great. Further, my own observation of intellectuals interested in politics would not confirm that *A* is high for them. They may have a great deal of information, but this seems to have been collected to confirm their basic position, not to change it.

Excerpted with revisions and permission from Gordon Tullock, *Toward a Mathematics of Politics* (Ann Arbor, Mich.: University of Michigan Press, 1972), pp.111-114

CHAPTER 17

International Trade and Finance

It can be of no consequence to America, whether the commodities she obtains in return for her own,, cost Europeans much, or little labor; all she is interested in, is that they shall cost her less labor by purchasing than by manufacturing them herself.

David Ricardo

Nations never really trade; people do. This simple point is important, for international trade allows us to approach international trade as an extension of models already developed, rather than a completely new topic. Earlier discussion focused on the local or national marketplace. In this chapter, our marketplace will be the world. We divide our discussion of international economics into its major subdivisions, *international trade* (mainly dealing with the exchange of real goods and services across national boundaries and their terms of trade) and *international finance* (mainly dealing with the exchange of national currencies and their exchange rates).

INTERNATIONAL TRADE

Of course, there are differences between international and domestic trade—enough to make international economics an important subdiscipline of the profession. Some differences are obvious, like the many different national currencies, cultures, institutions, laws, languages, artificial barriers (tariffs, quotas, embargoes, health regulations), and countercyclical domestic policies, involved in international exchange. Others go largely unrecognized. An intangible but significant factor is the difference in people's attitudes toward domestic and international trade—call international trade nationalism. As Abraham Lincoln is supposed to have said, "Domestic trade is among us; international trade is between us and them." Yet people all over the world trade with each other for the same reason: They stand to gain from the transaction in spite of the politics. There is much greater immobility of resources than commodities between nations. International trade is the substitute for the international movement of human and property resources, especially people.

Understanding that trade is between people, not nations, is important for another reason. If we focus solely on gains from trade to nations taken as unified political entities, we may overlook the distributional effects of international commerce—the gains and losses to individuals. As we will see, while international trade increases a nation's total income, international trade reduces some individual's incomes and increases others'. To evaluate objections to free trade among nations in proper perspective, we must recognize these hidden gains and losses.

Objections to free trade can be explained easily in terms of market theory. A major principle of economic theory is that each individual competitor has a vested interest in reducing competition. Competition forces product prices down and spurs product development and, in the long run, restricts business profits to only the risk-adjusted profit opportunities available elsewhere. Thus it is natural for domestic firms to seek protection from their foreign competitors—but protection only increases the prices consumers must pay. Carried to an extreme, protection based on the narrow interests of particular sectors of the economy can reduce everyone's income. On this basis rests the case for free international trade.

After examining the advantages of international trade from a purely national perspective, we will look at the distributional, or individual, effects. The chapter closes with a discussion of the pros and cons of protectionism.

Collective Gains from Trade

Most of the gains from trade result from allocating resources in the most efficient manner and from the reduction in the social opportunity cost—each geographic area produces and exchanges those things for which it is best suited to produce. With nations selling those things with the lowest opportunity costs, joint output is maximized and consumption opportunities are enhanced. Adam Smith told us more than two hundred years ago about the nature of the gains from trade: It is a maxim of every prudent master, never to attempt to make at home what it will cost him more to make than to buy." [1] Trade also allows a greater variety and wider choice of available products. The gains from it are clearest when there is no domestic substitute for an imported good. For example, the United States does not have any known reserves of chromium, manganese, or tin. For those basic resources, which are widely used in manufacturing, American firms must rely on foreign suppliers. The gains from trade are also clear for goods that are very costly or difficult to produce in the United States. For example, cocoa and coffee can be raised in the United States, but only in a greenhouse. Obviously it is less costly to import coffee in exchange for some other good, like wheat, for which the United States climate is better suited.

Foreign competition also offers benefits to the American consumer. By challenging the market power of domestic firms, foreign producers who market their goods in the United States reduce product prices and expand domestic consumption. Foreign competition also increases the variety of goods available. Without competition from the twenty or more foreign automobile producers who sell in the American market, the three U.S. automakers would each get a much larger percentage of the market. They would be loess hesitant to raise their prices if consumers had fewer alternative sources of supply. Collusion among major manufacturers would also be much more likely without the presence of foreign competitors.

International trade also promotes specialization, whose benefits are fairly clear. By concentrating on producing a small number of goods and selling to the world market, a nation can reap the benefits of greater efficiency and economies of scale. Resource

[1] Adam Smith, The Wealth of Nations (New York: Random House, Modern Library edition, 1937), p. 422.

savings that are not initially obvious may be gained. Indeed, after considering the following example, some readers may doubt that international trade can be mutually beneficial.

Consider a world in which only two nations, the United States and Japan, produce only two goods, textiles and beef. Assume that the United States produces both textile and beef more efficiently than Japan. That is, with the same resources, the United States can produce more beef and more textiles than Japan can. It has an absolute advantage in the production of both goods. An **absolute advantage** in production is the capacity to produce more units of output than a competitor can for any given level of resource use. A **comparative advantage** in production or cost is the relative advantage based on comparative ratios such that either the absolute advantage is greatest or the absolute disadvantage is smallest. Comparative advantage is more important for trade than absolute advantage. As long as the relative productivities or costs differ between individuals, regions, or nations, the participants can engage in mutually beneficial trade. Let's see how these differences work out for people.

Suppose that Lisa is worth $100 an hour in market work and only $10 an hour in home or household work. Her husband Gary is worth $8 an hour in the market and $4 in the home. Lisa has an absolute advantage in both tasks, but a comparative advantage in market work. She is ten times more productive in the market than at home; he is only twice as productive. Her comparative advantage (largest advantage is in the market; his comparative advantage (smallest disadvantage) is in the home. She should work in the market; and he should work at home. Their combined productivity would be $104 per hour (her $100 market rate plus his $4 home rate). If instead Gary worked in the market and Lisa worked at home, their combined productivity would be $18 (his $8 market rate plus her $10 home rate). They would be $86 (equal to $104 -- $18) better off by utilizing their comparative advantage, with Lisa working in the market, where her comparative advantage lies (her greatest absolute advantage, $92 over his) and Gary working at home, where his comparative advantage lies (his absolute disadvantage is smallest, $6 less than hers).

Table 17.1 shows these absolute and comparative differences for nations. With the same labor, capital, or other resources, the United States can produce thirty units of textiles; Japan can produce twenty-five. If the same resources are applied to beef production, the United States still outproduces Japan, by ninety units to twenty-five. Under such conditions, one might think that trade with Japan could not possibly benefit the United States. The relevant question is not how efficient the United States is in absolute terms, however, but whether the people of the United States can make a better deal by trading with Japan than they can make by trading among themselves.

This is determined by examining the comparative advantage, or the ratios of advantage or differences in relative efficiencies. A nation has a comparative advantage where (1) its absolute advantage is greatest or (2) its absolute disadvantage is smallest. Generally, a nation will have a comparative advantage in those products that require in their production a large proportion of factors that are relatively abundant and inexpensive in that nation and a comparative disadvantage in those productions that are relatively scarce and expensive in that nation. It is a technological fact that different products generally require in their production different proportions of the factors.

TABLE 17.1 Comparative Cost Advantages, Beef and Textiles, United States and Japan

	Maximum Units of Textiles (Zero Beef Units)	Maximum Units of Beef (Zero Textile Units)	Domestic Cost Ratios In Each Nation	Mutually Beneficial Trade Ratio, Both Nations
United States	30	90	1 textile costs 3 beef	1 textile trades for 2 beef
Japan	25	25	1 textile costs 1 beef	

To determine which is the better deal, we must compare the costs of production. We know that there is an uneven distribution of economic resources among nations. This produces differences in productive capacities based on these differences in relative factor endowments. If each nation produces and trades the products in which it has a comparative cost advantage, trade can raise both their incomes. Remember that a comparative advantage is the capacity to produce a product at a lower cost than a competitor, in terms of the goods that must be given up. The United States may have an absolute advantage in the production of both beef and textiles, but it may have a comparative advantage only in the production of beef. In other words, the United States must forgo fewer units of textiles to obtain a unit of beef than Japan. Although a single nation could theoretically have an absolute advantage in all commodities, it could not have a comparative advantage in all commodities. With two nations and two commodities, *if a nation has a comparative advantage in one commodity it must have a comparative disadvantage in the other commodity.* Having a comparative advantage in beef necessarily means the United States cannot have a comparative advantage in textiles—a point that will become clear shortly.

In a sense, the United States trades with itself every time it producers either beef or textiles. If it produces beef, it incurs an opportunity cost; it gives up some of the textiles it could have produced. If it produces textiles, it gives up some beef. In Table 17.1, every time the United States produces one unit of textiles, it gives up three units of beef. (It can produce either thirty units of textiles or ninety of beef—a ratio of one to three.) Thus the United States can benefit by trading beef for textile if it can give up fewer than three units of beef for each unit of textiles it gets from Japan.

Japan, on the other hand, gives up an advantage of one unit of beef for each unit of textiles it produces. If Japan can get more than one unit of beef for each unit of textiles it trades, it too can gain by trading. In short, if the trade ratio is greater than one unit of beef for one unit of textiles but less than three units of beef for one unit of textiles, trade will benefit both countries. The United States will gain because it has to give up fewer units of beef—two, perhaps, instead of three—than if tried to produce the textiles

itself. It can produce three units of beef, trade two of them for a textile unit, and have one extra beef unit left over—or it can trade all three units of beef for one and one-half units of textiles. Japan can produce one unit of textiles and trade it for two units of beef, gaining one textile unit in the process.

Both nations can gain from such a trade because each is specializing in the production of a good for which it has a comparative opportunity cost advantage.[2] Even though the United States has an absolute cost advantage in both products, Japan has a comparative advantage in textiles. One unit of textiles costs Japan one unit of beef; the same unit of textiles costs the United States three units of beef. Similarly, the United States has a comparative cost advantage in the production of beef. One unit of beef costs the United States only one-third unit of textiles; it costs Japan a whole unit. If each country specializes in the commodities for which it has a comparative cost advantage, the two nations can save resources for use in further production.

TABLE 17.2 Mutual Gains from Trade in Beef and Textiles, United States and Japan

	United States	**Japan**	**Total, U.S. and Japan**
Production and consumption levels before international trade	15 textiles 45 beef	3 textiles 22 beef	18 textiles 67 beef
Production levels in anticipation of international trade (complete specialization assumed)	0 textiles 90 beef	25 textiles 0 beef	25 textiles 90 beef
At an exchange ratio of 2 beef for 1 textile, United States and Japan agree to trade 40 beef for 20 textiles.			
Consumption levels after international trade	20 textiles 50 beef	5 textiles 40 beef	25 textiles 90 beef
Increased consumption (before-trade consumption levels subtracted)	5 textiles 5 beef	2 textiles 18 beef	7 textiles 23 beef

Table 17.2 shows the gains in production each nation can realize under such an arrangement. Before trade, the United States produces 15 units of textiles and 45 of beef; Japan produces 3 units of textiles and 22 of beef. Total production is therefore 18 units of textiles and 67 units of beef. With trade, the United States produces 90 units of beef

[2] Specialization in production for the United States and Japan will likely be partial with increasing marginal production costs. With constant-cost or decreasing-cost, the specialization of production may be complete.

and Japan produces 25 units of textiles. At an international trade ratio of 1 unit of textiles to 2 units of beef, suppose the two nations agree to trade 40 units of beef for 20 units of textiles. The United States gets more beef—50 units as opposed to 45—and more textiles—20 units as opposed to 15. Japan also gets more of both commodities. Through specialization, total world production has risen from 18 to 25 units of textiles and from 67 to 90 units of beef. Both nations can now consume more of both commodities. In a very important sense, the world's aggregate real income has increased.

The same gain in aggregate welfare is shown graphically in Figure 17.1. On the left side of the figure, the U.S. production possibilities curve extends from 30 units of textiles on the horizontal axis to 90 units of beef on the vertical axis. Japan's production capability is shown on the right. Without trade, the United States chooses to produce at point *a*, 15 textiles units and 45 beef units. At an exchange ratio of 2 beef units for one textile unit, the United States can move up and to the left on its production possibilities curve. At the extreme, it will produce at point *b*, 90 units of beef and no textiles. It can trade along the outer line, exchanging 40 beef units for 20 textile units (point *c*). Through trade, the United States realizes a gain in aggregate welfare represented by the distance

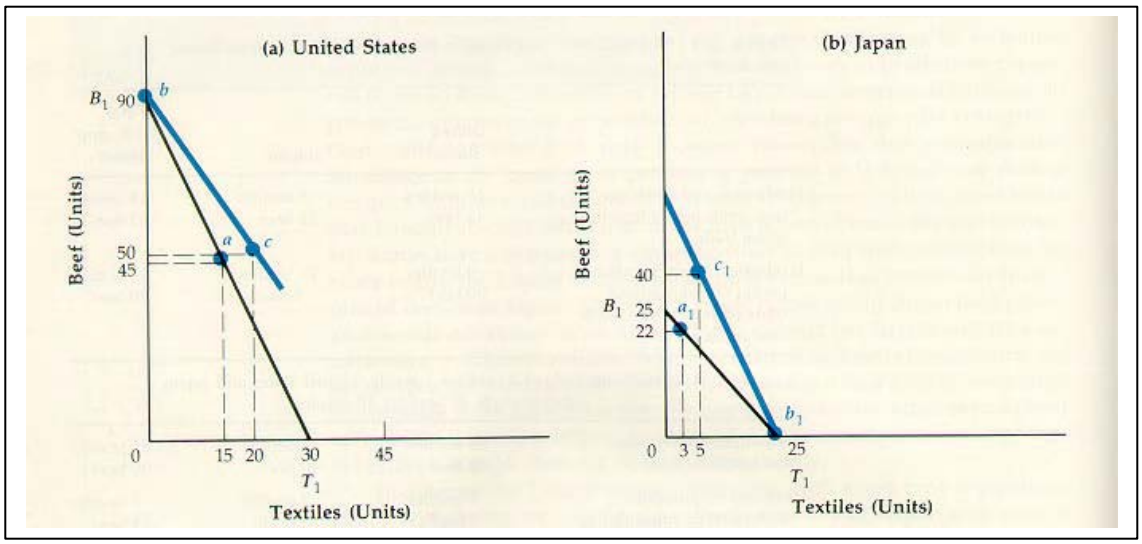


FIGURE 17.1 Production Gains from International Trade

The United States can produce any combination of beef and textiles along its production possibilities curve B_1T_1 (left panel). Without trade, it will choose to produce at point *a*, 45 units of beef and 15 units of textiles. If given the opportunity to trade two units of beef for one unit of textiles, however, the United States will specialize completely in beef (point *b*) and trade beef for textiles along the darkened line. Through trade, the United States moves from *a* to *c*, exporting 40 units of beef (90 units produced minus 50 consumed) and importing 20 units of textiles. In the process the nation increases its consumption of both beef and textiles, from 45 units of beef and 15 units of textiles to 50 units of beef and 20 units of textiles. (the darkened line does not intersect the horizontal beef axis because the United States cannot get more than 25 units of textiles from Japan.). At the same time trade permits Japan (right panel) to shift its consumption from the black production possibilities curve to the darkened curve. By producing at b_1 and exporting 20 units of textiles in exchange for 40 units of beef, Japan too can expand its consumption, from a_1 to c_1.

between points *a* and *c*. In other words, international trade permits the United States to consume at a point beyond its own limited production possibilities curve (the black line in the graph). In the same way, Japan realizes a gain in welfare equal to the difference between its consumption before trade, a_1, and its consumption after trade, c_1.

In the long run, a country's imports are paid for by its exports. Thus, by engaging in international trade, according to comparative advantage, a country gains by reducing its social opportunity cost. The social opportunity cost of imports is the exports required to pay for the imports. If the resources used to produce exports are less than those required to produce the goods domestically, there is a net social economic gain.

The Distributional Effects of Trade

As we have seen, even a nation that has an absolute advantage in every production process can benefit from trade. In reality, no such nation exists, but that just underscores the point that even in the unlikeliest of conditions, we can make the case for free trade. Furthermore, if voluntary trade takes place we must assume that both parties perceive that they will gain. Why else would they agree to the arrangement? How much each nation gains depends on the **terms of trade**—the ratio at which one commodity can be traded or exchanged for another commodity internationally; or on an aggregate basis, it is the ratio of the price of exports to the price of imports. The more favorable a nation's terms of trade, and therefore its exchange rate, the larger its share of gain in enhanced output.

International trade remains a controversial subject, for although nations gain from trade, individuals within those nations may not. Individual gains tend to go to the firms that produce goods and services for export, losses tend to go to the firms that produce goods and services that are imported under free trade.

Gains to Exporters

Exporters of domestic goods gain from international trade because the market for their goods expands, increasing demand for their products. The increase in their revenue can be seen in Figure 17.2. When the demand curve shifts from D_1 to D_2, producers' revenues rise from P_1Q_2 (point *a*) to P_2Q_3 (point *b*). The more price-elastic or flatter the supply function (*S*), the larger the change in quantity and the smaller the change in price. The increase in revenues is equal to the shaded L-shaped area $P_2Q_3Q_2aP_1$. Producers benefit because they receive greater profits, equal to the shaded area above the supply curve, P_2baP_1. Workers and suppliers of raw materials benefit because their services are in greater demand, and therefore more costly. The cost of producing additional units for export is equal to the shaded area below the supply curve, Q_2abQ_3.

This graph suggests why farmers supported the sales of wheat to the Soviet Union that began in the early 1970s. They complained loudly when the U.S. government suspended sales temporarily for political reasons. Many consumers and members of Congress objected the wheat sales, however, on the grounds that they would increase the domestic price of wheat and therefore of bread. In a narrow sense, consumers of

exported products have an interest in restricting their exportation. Yet in the broad context of international trade. Restrictions can work against the private interests of individuals, including even consumers of bread. Trade is ultimately a two-way street. To import goods and services that can be produced more cheaply abroad than at home, a nation must export something else. No nation will continually export part of what it produces without getting something in return. To the extent that exports are restricted to suit the special interests of some group, imports of other commodities also are restricted. Restrictions on the exportation of wheat may hold down the price of bread, but they can also increase the price of imported goods, like radios and television sets.

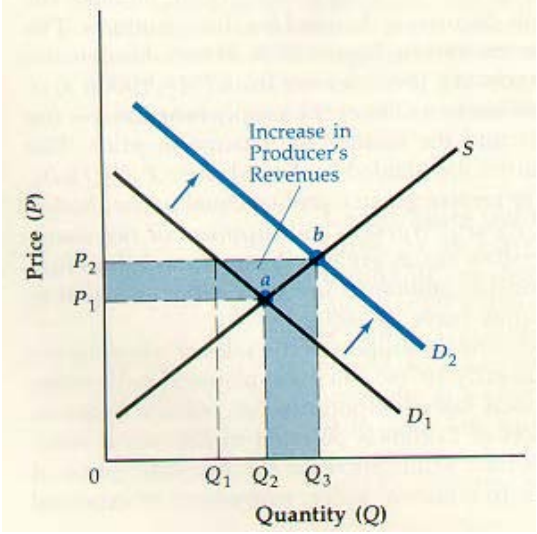


FIGURE 17.2 Gains from the Export Trade

The opening up of foreign markets to U.S. producers increases the demand for their products, from D_1 to D_2. As a result, domestic producers can raise their price from P_1 to P_2 and sell a larger quantity, Q_3 instead of Q_2. Revenues increase by the shaded area $P_2bQ_3Q_2aP_1$. The more price-elastic or flatter the supply function (S), the larger the change in quantity and the smaller the change in price.

Losses to Firms Competing with Imports

While consumers gain from increased imports, domestic producers may lose from increased competition. Foreign producers can gain a foothold in the domestic market in three ways: (1) by providing a better product than domestic firms; (2) by selling essentially the same product as domestic firms, but at a lower price; and (3) by providing a product previously unavailable in the domestic market. Most people welcome the importation of a previously unavailable product, but producers who face competition from foreign suppliers have an incentive to object to importation. If imports are allowed, the domestic supply of a good increases. Domestic competitors will sell less, and they may have to sell at a lower price. In short, the employment opportunities and real income of domestic producers decline as a result of foreign competition.

Figure 17.3 shows the effects of importing foreign textiles. Without imports, demand is D and supply is S_1. In a competitive market, producers will sell Q_2 units at a price of P_2. Total receipts will beP_2 x Q_2. The importation of foreign textiles increases the supply to S_2, dropping the price from P_2 to P_1. Because prices are lower, consumers increase their consumption from Q_2 to Q_3 and get more for their money. The more price-elastic or flatter the demand curve (D), the greater the change in quantity and the smaller the change in price.

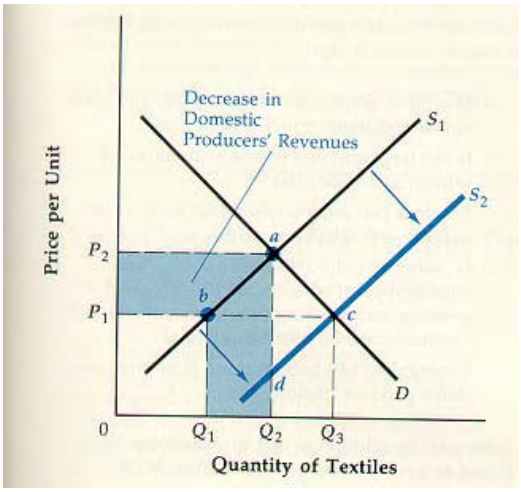


FIGURE 17.3 Losses from Competition with Imported Products

The opening up of the market to foreign trade increases the supply of textiles from S_1 to S_2. As a result, the price of textiles falls from P_2 to P_1, and domestic producers sell a lower quantity, Q_1 instead of Q_2. Consumers benefit from the lower price and the higher quantity of textiles they are able to buy, but domestic producers, workers and suppliers lose. Producers' revenues drop by an amount equal to the shaded area P_2abP_1. Workers' and suppliers' payments drop by an amount equal to the shaded area Q_2abQ_1. Starting at point *c*, a tariff or tax equal to *ad* is levied, shifting the supply curve from S_2, S_1. In an industry whose costs are increasing, the increase in price from P_1 to P_2 in the importing country is less than the increase in the tariff (*ad*), because a price fall in the exporting country absorbs some of the burden of the duty.

Domestic firms, their employees, and their suppliers lose. Because the price is lower, domestic producers must move down their supply curve (S_1) to the lower quantity Q_1. Their revenues fall from P_2Q_2 to P_1Q_1. In other words, the revenues in the shaded L-shaped area $P_2a\,Q_2Q_1bP_1$ are lost. Of this total loss in revenues, owners of domestic firms lose the area above the supply curve, P_2abP_1. Workers and suppliers of raw materials lose the area below the supply curve, Q_2abQ_1. This is the cost domestic firms would incur in increasing production from Q_1 to Q_2, the payments that would be made to domestic workers and suppliers in the absence of foreign competition. If workers and other resources are employed in textiles because it is their best possible employment, the introduction of foreign products can be seen as a restriction on some workers' employment opportunities. In summary, while international trade lowers import prices and raises export prices in the domestic nation, the net impact is a reduced social opportunity cost curve that expands total output and consumption opportunities.

The Effects of Trade Restrictions Such as Tariffs and Quotas

Because foreign competition hurts some individuals, domestic producers, workers, and suppliers have an incentive to seek government restrictions on the imports of tradables. Of course, some industries such as communications, services, and utilities are largely insulated from foreign competition without trade restrictions. Two forms of protection are commonly used, tariffs and quotas. A **tariff** is a special tax or duty on imported goods that can be a percentage of the price (*ad valorem* duty) or a specific amount per unit of the product (specific duty). A tariff may be imposed to raise money for the

levying country—typically, revenues are modest on commodities not produced in the levying country—or in the more likely case, to protect some industry against the cold winds of competition. A **quota** is a physical or dollar value limit—mandatory or voluntary—on the amount of a good that can be imported or exported during some specified period of time. There are other nontariff barriers such as controlling the flow of foreign exchange, licensing requirements, health, quality, or safety restriction and regulations on products.

If tariffs are imposed on a foreign good such as textiles, the supply of textiles will decrease—say, from S_2 to S_1 in Figure 17.3—and the price of imports will rise. Domestic producers will raise their prices too, and domestic production will go up. If the tariff is high and all foreign textiles are excluded, the supply will shift all the way back to S_1. A tariff will have a more modest effect, shifting the supply curve only part of the way back toward S_1. The price of textiles will rise and domestic producers will expand their production, but imports will continue to come into the country. How much the price rises and the quantity falls after the imposition of the tariff depends on how price-elastic or flat the demand curve (D) is. The more elastic D is, the greater the fall in quantity and the greater the rise in price. The imposition of a duty can cause the taxed good in the importing country to increase by exactly the amount of the duty, less than the duty, or in the extreme case, not at all (depending on price elasticity). In the most likely case (of increasing cost conditions and a rising supply curve) a tariff will cause the price to increase in the importing country by less than the amount of the duty as the price falls in the foreign country. The tariff will cause the domestic and foreign price to differ by exactly the amount of the tariff, but the price increase in the importing country is equal to the tariff minus fall in price in the exporting country. Thus, in Figure 17.3, starting from point c, the increase in the price in the importing country from P_1 to P_2 is less than the tariff equal to ad, shifting the supply curve from S_2 to S_1 as part of the duty is shifted to the exporting country where the price falls. For instance, a tariff of $3 per unit may cause the import price to rise by $2 and the export price to fall by $1 with both nations absorbing part of the burden of the tariff. Who bears the biggest burden is a matter of relative price elasticity, just as whether buyers or sellers bear the burden of a domestic excise tax. As always, the more inelastic the demand of the buyers and the more elastic the supply of the sellers, the bigger burden of any tax—domestic (e.g., excise) or foreign (e.g., an import duty)—that falls on the buyers.

A quota has the same general effect as a tariff, although its price-cost effect can be much more drastic. They both reduce the market supply, raise the market price, a encourage domestic production, thereby helping domestic producers and harming domestic consumers. A quota, however, can sever international price-cost links because the market mechanism for relating the prices of different nations is artificially stopped from functioning. Nonetheless, quotas are sometimes imposed by nations because they are a more certain and precise technique of control, and can be changed by administrative decree.

There are three main differences between quotas and tariffs. First quotas firmly restrict the amount of a product that can be imported, regardless of market conditions. A quota may specify how much oil may be imported each day or how much sugar may be imported each year. Tariffs, on the other hand, permit any level of importation for which

consumers are willing to pay. Thus, if demand for the product increases, imports may rise. (There is a hybrid called a "tariff quota" that sets a fixed limit on importation or exportation.)

The first Reagan administration imposed quotas on steel, copper, textiles, and autos from Japan. In 1984 the so-called voluntary restraint program forced Japan to restrict auto sales in the United States to 1.84 million cars. Foreign cars now represent about 25 percent of U.S. sales. Because Japanese supply was not allowed to keep pace with the rapidly expanding U.S. demand, the price of Japanese cars rose, more expensive models were imported, and consumers faced longer waiting lists for Japanese cars. The price of American cars also rose. These consequences led to the termination of the voluntary restraint program in 1985.

The second major difference between tariffs and quotas is that quotas are typically specified for each important foreign producer. Otherwise, all foreign producers would rush to sell their goods before the quota was reached. When quotas are rationed in this way, more detailed government enforcement is required. Tariffs place no such restrictions on individual producers. Moreover, the tariff is collected by the government in custom duties while price enhancement with a quota goes as a windfall gain to the fortunate few with import licenses.

Finally, quotas enable foreign firms to raise their prices and extract more income from consumers. One economist estimated that the Reagan administration's voluntary restraint program permitted Japanese auto producers to raise their prices high enough to take an additional $2,500 per car, or $5 billion, out of the American market.[3] As a result of the protectionist shield, U.S. automakers raised domestic car prices $1,000 per car, or $8 billion per year, in 1984 and 1985. Tariffs, on the other hand, force foreign firms to lower their prices to offset the increase from the tariff. They also generate income for the federal government. Although tariffs and quotas promote a less efficient allocation of the world's scarce resources, because of the private benefits to be gained from tariffs and quota, we should expect an industry to seek them as long as their market benefits exceed their political cost. Politicians are likely to expect votes and campaign contributions in return for tariff legislation that generates highly visible benefits to special interests. Producers (and labor) will usually make the necessary contributions, because the elimination of foreign competition promises increased revenues in the protected industries. The difference between the increase in profits due to import restrictions and the amount spent on political activity can be seen as a kind of profit in itself. Surprisingly, protectionism may sometimes also be supported by exporters, as a tariff or quota can stimulate net exports. Since protectionism also causes the exchange rate to appreciate, however, this discourages exports and offsets partially or wholly the tariff-driven increase in net exports.

Consumers, on the other hand, have reason to oppose tariffs or quotas on imported products. Such legislation inevitably causes prices to rise, as a tariff amounts to a subsidy to the domestic producer of the dutiable product, paid for largely by the consumers of that product in the form of higher prices. Consumers typically do not offer

---

[3] Robert Crandell, "Assessing the Impact of the Automobile Export Restraints upon U.S. Automobile Prices," mimeo, Brookings Institution, December 1985.

very much resistance, however, because the effects of tariffs and quotas are hard to perceive. Unlike a sales tax, the cost of a tariff is not rung up separately at the cash register, and many consumers do not reason through the complex effects of a tariff on consumer prices. In fact, many if not most, consumers feel that tariffs on foreign automobile, steel, or copper producers are good for the nation and for themselves. "Buy American'' slogans and advertisements emphasizing the need to preserve American jobs are generally effective in swaying public opinion. One comprehensive investigation showed that protection in thirty-one countries cost consumers $53 billion in 1984, while providing only $40 billion in benefits to the producers.[4]

As a group, consumers have less incentive to oppose tariffs than industry has to support them, as the costs to individual consumers and taxpayers are negligible and largely hidden. The benefits of a tariff accrue principally to a relatively small group of firms, whose lobby may already be well entrenched in Washington. These firms have a strong incentive to be fully informed on the issue and to make campaign contributions, but the harmful effects of a tariff are diffused over an extremely large group of consumers. The financial burden any one consumer bears may be very slight, particularly if the tariff in question is small, as most tariffs are. As result, the individual consumer has little incentive to become informed on tariff legislation or to make political contributions to lobbies that support such legislation. Although consumers as a whole may share an interest in opposing tariffs, collective action must still be undertaken by individuals—and individuals will not incur the cost of organizing unless they expect to receive compensating private benefits.

At some level of increased cost, of course, consumers will find the necessary incentive to oppose tariff legislation. For this reason Congress rarely passes tariffs high enough to make importation totally unprofitable. Even low tariffs reduce the nation's real income while redistributing it toward protected sectors. The size of the pie is reduced, but the protected few get a bigger slice. In spite of all the impediments to free trade imposed by U.S. economy, there has been a substantial increased in our dollar volume of imports and exports over the last thirty years. Similarly, world trade has increase in the last three decades. Over 15 percent of the world's production is now consumed in a different nation than where it was produced. Put differently, the dollar value of imports to all countries has increased tenfold since 1960.

According to Alan S. Blinder,[5] the case against protectionism, described as a negative-sum game, where the losing consumers lose more than the winning protected producers win, involves even more problems. There are four other problems with trade restrictions. First, protectionism allows high-cost producers that would otherwise fail to survive. Second, trade restrictions have a habit of affecting other industries. For example, automobiles need protection because the ball bearings, steel, and textiles that provide inputs to automobiles are protected. Third, foreign nations often retaliate against protectionism. Tit-for-tat is the modus operandi in international trade: Country A raises barriers on product X because Country B did it to product Y. Fourth, trade restrictions

---

[4] Gary C. Hufbauer, et al., *Trade Protection in the United States: 31 Case Studies* (Washington, D.C.: Institute for International Economics 1986).

[5] Alan Blinder, *Hard Heads, Soft Hearts* (Reading, Mass.: Addison-Wesley, 1987), pp. 118-119.

aren't really job-saving or job-creating, but job-swapping. Protectionism raises the exchange rate, hurting exports in unprotected industries. Because in the long run the value of exports must be equal to the value of imports, we end up swapping jobs in efficient unprotected industries.

The Case for Free Trade

We have seen how international trade can on balance increase the total incomes of the nations engaged in it, although export producers gain and import-substitute producers lose. By extension, we can conclude that anything that restricts the scope of trade between nations generally reduces their real incomes. To the extent that trade is a two-way street—that exports trade for imports, at least in the long run—a reduction in imports brings a reduction in exports. From our imports the Japanese get the dollars they need to buy American exports. If we reduce our imports, they will have fewer funds with which to buy from us. For this reason, U.S. farmers, who sell approximately one-third of their crops in foreign markets, actively opposed the protectionist movement led by textiles, steel, and copper firms in the 1980s.

Yet what is true for one sector of the economy is not necessarily true for all. If all sectors are protected by tariffs, it is possible (but not inevitable) that all experience a drop in real income. Figure 17.4 illustrates the case of an economy with two industries, automobiles and textiles. Both industries must compete with imports. If neither seeks protection, both will operate in cell I, at a combined real income of $50 ($20 for the textiles industry and $30 for the automobile industry). If the textiles industry seeks protection but the auto industry does not, they will move to cell II, where tariffs raise the textiles industry's income from $20 to $23. The automotive sector's income falls to $25, so that the two industries' combined real income falls to $48. Consumers get fewer textiles at a higher price.

Similarly, if the auto industry seeks protection while the textiles industry does not, the economy will move from cell I to cell III. Again, total real income falls from $50 to $49, but this time the auto industry is better off. Its income rises from $30 to $34, while the textiles industry's income falls to 15. Obviously, if one industry seeks protection, the other has an incentive to follow suit. If the textiles industry counters with a tariff of its own, the economy will move from cell III to cell IV, and the industry's real income will rise from $15 to $17.

Without some constraint on both sectors, then, each has an interest in seeking protection regardless of what the other does. Yet if the economy winds up in cell IV, total real income will be lower than under any other conditions: only $43. Obviously the best course for the economy as a whole is to prohibit tariffs altogether, and in an economy with only two sectors, the cost of reaching an agreement is manageable. In the real world, however, there are many economic sectors, and the costs of reaching a decision are much greater.

In Figure 17.4, both industries end up with lower real incomes in cell IV, but in reality, the effects of multiple tariffs will be different in different sectors of the economy. Although total real income will fall, several sectors may realize individual gains.

Consider Figure 17.5. Although total real income falls from cell I ($50) to cell IV ($48), the auto sector's income rises (from $30 to $31). In this case the textile sector bears the brunt of tariff protection, and the auto sector has a compelling interest in obtaining protective tariffs. The sectors of the economy that are most adept at manipulating the political process will be the least willing to accept free trade.

Although it is true that for a nation some trade is better than no trade, it is not necessarily true that free trade is better than restricted trade. Even though protectionism promotes economic inefficiency in the aggregate, a nation may under certain conditions act like a monopolist and improve its share of the gains through trade restrictions. Similarly, the owners of relatively scarce factors of production may be better off with little or no trade.

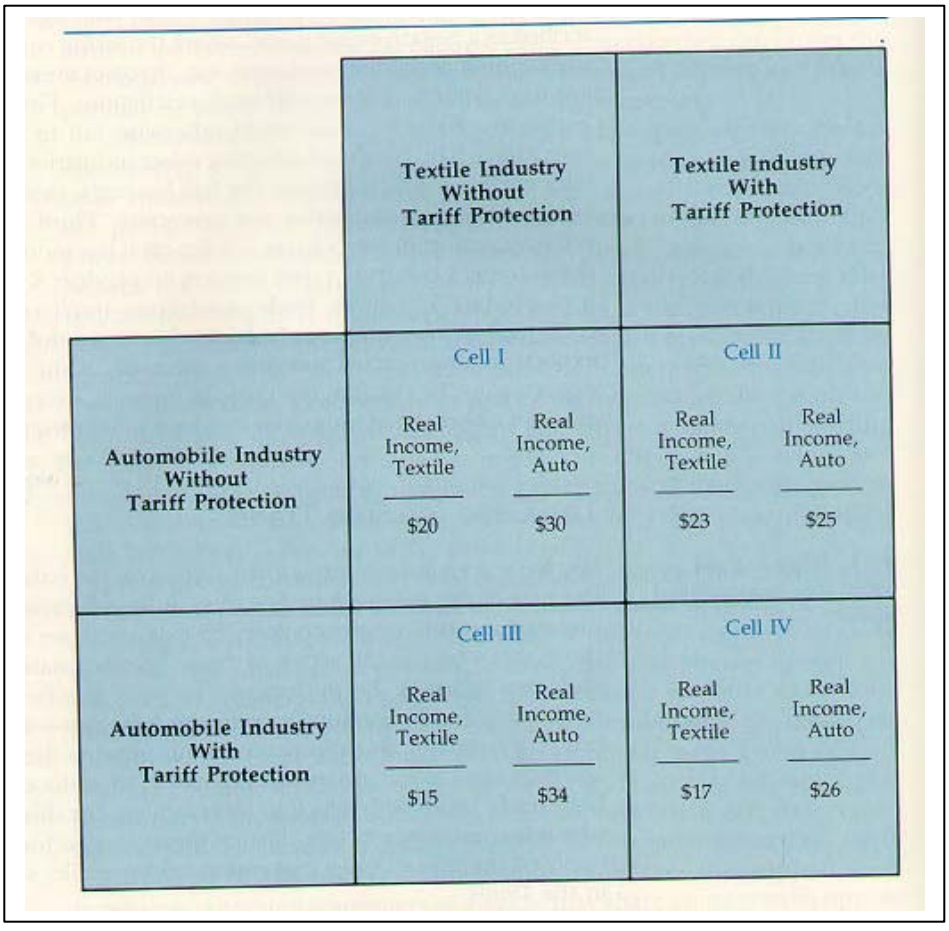


FIGURE 17.4 Effects of Tariff Protection on Individual Industries: Case 1

If neither the textiles nor the automobiles industry obtains tariff protection, the economy will earn its highest possible collective income (cell I), but each industry has an incentive to obtain tariff protection for itself. If the textiles industry alone seeks protection (cell II), its income will rise while the auto industry's income falls. If the auto industry alone seeks protection, its income will rise while the incomes of textiles income industry falls. If both obtain protection, the economy will end up in cell IV, its worst possible position. Income in both sectors will fall.

Thus the case for free trade is a subtle one. As always, special-interest group -- entrepreneurs, labor organizations, consumer groups -- will pursue their individual interests, competing for favors and benefits the same way they compete in the marketplace. Yet if all are to be treated equally by government, we must make the choice between free trade for all and protection for all.

Economists generally choose free trade for all, because of its obvious benefits to the nation as a whole. There are some legitimate exceptions to that rule, such as the required domestic production of public goods, which are discussed below. Yet even trade restrictions necessary for the public good are abused by those who would secure protection for private purposes.

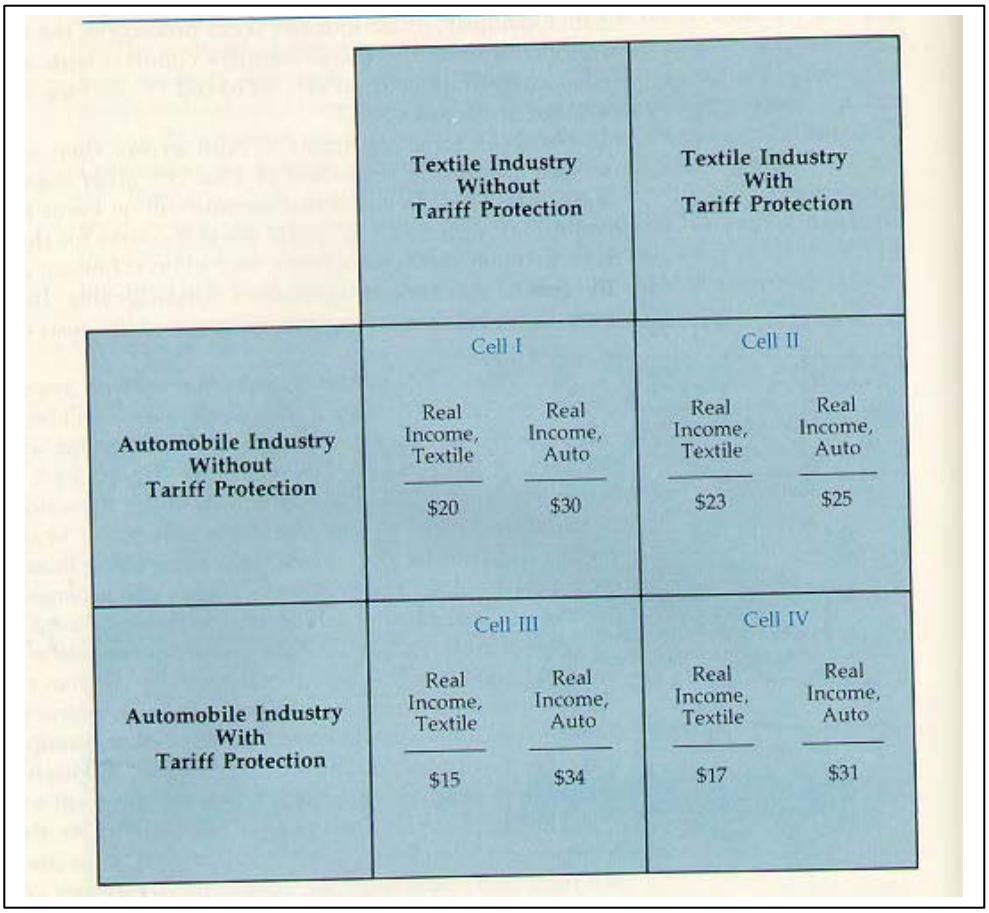


FIGURE 17.5 Effects of Tariff Protection on Individual Industries: Case 2

In this more realistic case, the auto industry gains from tariff protection, even if both sectors are protected (cell IV). The textiles industry's income falls from $20 (cell I) to $17 (cell IV), but the auto industry's income rises from $30 (cell I) to $31 (cell IV). Thus the auto industry has no incentive to agree to the elimination of tariffs.

The Case for Restricted Trade

Proponents of tariffs rarely argue publicly that they will serve private interests, raise prices, and reduce the availability of goods. Instead, they typically advocate tariffs as the most efficient means to accomplishing some national objective. Any private benefits that would accrue to protected industries are generally portrayed as insignificant side effects.

Although most arguments in favor of tariffs camouflage the underlying issues, one is partially valid. It has to do with the maintenance of national security.

The Need for National Security

Protariff arguments based on national or military security stress the need for a strong defense industry. If imports are completely unrestricted, certain industries needed in time of war or other national emergency could be undersold and run out of business by foreign competitors. In an emergency, the United States would then be dependent on possibly hostile foreign suppliers for essential defense equipment. (The nation could convert to production of war-related goods, but the conversion process might be prohibitively lengthy and complex.) Tariffs may create inefficiencies in the allocation of world resources, but that is one of the costs a nation must bear to maintain military self-sufficiency and hence a strong national defense.

Given the unsteady popularity of U.S. foreign policy and the uncertain support of allies, this argument has some merit. Other nations, like Israel, have found that they cannot count on the support of all their allies in time of war. Because France disagreed with Israeli policy in the Middle East, it held up shipment of spare parts for planes it had sold to Israel earlier. The United States could conceivably find itself in a similar position if it relies on foreign firms for planes, firearms, and oil.

Special-interest groups can easily abuse the national defense argument for tariffs. The textile industry, for example, promotes itself as a ready source of combat uniforms during wartime. Even candle manufacturers have petitioned Congress for increased tariff protection, on the grounds that candles are "a product required in the national defense."[6] In years past, U.S. oil producers, contending that a healthy domestic oil industry is vital to the national defense, have lobbied for protection from foreign oil in wartime, the effects of a tariff are not entirely straightforward as might be thought. By making foreign oil more expensive, a tariff increases consumption of domestic oil. Since oil is a finite resource, a tariff can ultimately make the United States more dependent on foreign energy sources in time of emergency.

Recent history illustrates the danger of dependence on foreign suppliers. In 1973, the OPEC oil cartel used U.S. dependence on its oil reserves as a bargaining tool in its efforts to reduce U.S. support for Israel. President Gerald Ford responded in 1974 by supporting a tariff on imported oil, to stimulate exploration for new domestic energy reserves. If the United States could become energy, independent by the end of the 1980s, Ford argued, it would reduce the threat of political blackmail from the Middle East. In

---

[6] "Petition of the Candlemakers—1951," in *Readings in Economics*, ed. Paul Samuelson (New York: McGraw-Hill, 1973), 7th ed., p. 237.

1983, for the same reason, the Reagan administration granted tariff protection to specialty steel products, which are used extensively in high-technology defense systems.

Other Arguments

Most of the other arguments in support of tariffs are weak from a practical as well as a theoretical perspective. In fact, while protectionism is a growth industry in recent years, the costs to society exceed the benefits. It is sometimes argued that because workers are paid less in foreign countries, U.S. industries cannot hope to compete with foreign imports—but trade depends on the relative costs of production, not absolute wage rates in various nations. U.S. wages may be quite high in either absolute or relative terms. If U.S. workers are more productive than others, however, the costs of production can be lower in the United States than elsewhere.

The important point is what tariffs do to trade. In an earlier example of trade in textiles and beef, the United States was more efficient than Japan in the production of both products. That is, generally speaking, fewer resources were required to produce those goods in the United States than in Japan. Very possibly, the incomes of textiles and beef workers would be higher in the United States than in Japan, but because Japanese firms had a comparative cost advantage in textiles (measured in terms of the number of units of beef forgone for each textiles unit), they were able to undersell textiles firms in the United States. If the U.S. imposed tariffs or quotas on imported textiles because Japan had a comparative advantage in that product, it would destroy the basis for trade between the two nations. Reducing imports will tend to reduce exports, at least in the long run.

A second questionable argument for tariffs is based on the faulty idea that the United States loses when money flows overseas in payment for imports. As Abraham Lincoln is reported to have said, "I don't know much about the tariff, but this I do know. When we trade with other countries, we get the goods and they get the money. When we trade with ourselves, we get the goods and the money."

Lincoln was clearly right when he said he did not know much about the tariff. He failed to recognize the real income benefits of international trade, which are reduced by tariffs. He seems to have confused the nation's welfare with its monetary holdings. It is true that if Americans buy goods from abroad, they get the goods and foreigners get the money.[7] What are foreigners going to do with the money they receive, however? If they never spend it, Americans will be better off, for they will have gotten some foreign goods in exchange for some paper bills, which are relatively cheap to print. At some point, however, foreign exporters will want to get something concrete in return for their labor and materials. They will use their dollars to buy goods from U.S. manufacturers. Again, trade is a give-and-take process, in which benefits flow to both sides.

A third argument often made is that foreign nations impose tariffs on U.S. goods; unless we respond in kind, foreign producers will have the advantage in both markets.

---

[7] Actually, the transaction may not involve the transfer of paper money. It is more likely—as explained in the next chapter—that payment will be made by transferring funds from one bank account to another. The importer's bank balance will drop, and the exporter's bank balance will increase.

This argument has a significant flaw. By restricting their imports, foreign nations reduce their ability to sell to the United States and other nations. To buy Japanese goods, for instance, Americans need yen. They get yen by selling to Japan. If Japan reduces its imports from the United States, Americans will have fewer yen to buy Japanese goods. So the Japanese are restricting their own exports with their tariffs. They harm themselves as well as Americans. If Americans respond to their actions by imposing tariffs of their own, they will reduce trade even further. The harm is compounded, not negated.

One sound reason for increasing tariffs is to strengthen our bargaining position in international trade conferences. By matching foreign restrictions, the United States may be able to force a multilateral reduction of tariffs. To the extent that all tariffs are reduced by such a strategy, world trade will be stimulated.

According to the fourth argument, tariffs increase workers' employment opportunities. If the government imposes tariffs on imported goods, the demand for American goods will rise. More workers will have jobs and can spend their income on goods and services produced by other Americans.

It is true that in the short run, more workers are likely to be hired because of tariffs, but in the long run reduced imports will result in reduced exports. The market for U.S. goods will shrink, increasing unemployment in the export industries.

Furthermore, if Americans reduce their demand for foreign goods to increase employment in the United States, their domestic recession will be transmitted to other nations. With fewer sales of foreign goods, fewer workers will be needed in foreign industries. Foreign governments may retaliate by imposing tariffs of their own. Tariffs will temporarily increase their employment levels and can be used as a bargaining tool in trade negotiations as well. The end result will be a reduction in total worldwide production and real income.

Finally, tariff advocates sometimes claim that new industries deserve protection because they are too small to compete with established foreign firms. If protected by tariffs, these new industries can expand their scale of production, lower their production costs, and eventually compete with foreign producers.

It is very difficult, however, for a government to determine which new industries may eventually be able to compete with foreign rivals. Over the long period of time that an industry needs to mature, conditions, including the technology of production, may change significantly. For a so-called infant industry to become truly competitive, furthermore, it must develop a comparative cost advantage, not just economies of scale.

Moreover, the mere likelihood that a firm will eventually be able to compete with its foreign rivals does not in itself warrant protection. Not until firms have become established will consumers receive the benefit of lower prices. In the interim, tariff protection hurts consumers by raising the prices they must pay. Proponents of protection must be able to show that the time-discounted future benefits to be gained by establishing an industry exceed the current costs of protecting it.

Finally, if a firm can expand, cover all its costs of production, and eventually compete with it foreign rivals, private entrepreneurs are not likely to miss the opportunity to invest in it. Through the stock and bond markets, firms with growth potential will be

able to secure the funds they need for expansion. If a firm cannot raise capital from private sources, it may be because the return on the investment is too low in relation to the risk. Why should the government accept risks that the private market will not accept?

INTERNATIONAL FINANCE

People rarely use barter in trade. Exchanging one toy for two pens or three pots for the rear end of a steer simply is not practical. Because the bartering seller must also be a buyer, buyers and sellers may have to incur very substantial costs to find one another, even in the domestic market. When people are hundreds or thousands of miles apart and separated by national boundaries and foreign cultures and languages, as they are in international trade, barter would be all the more complicated. We rarely see exporters acting as importers, exchanging specific exports for specific imports.

In the domestic economy, money reduces the cost of making exchanges. The seller of pots needs only to find a buyer willing to pay with bills, coins, or a check. He does not have to accept goods that may be difficult to store, use, and trade. In the international economy too, money facilitates trade, but well over a hundred different national currencies are in use. The French have the franc; the Japanese, the yen; the Americans, the dollar. To deal with this complication, a system of international exchanges emerged in which importers pay for the goods they buy in their currency. Before international trade can take place, it is usually necessary for the country buying to convert to the currency of the trading partner. Importers demand foreign currency and exporters supply it. How the international monetary system works, and the problems inherent in it, are the subjects of this section.

The Process of International Monetary Exchange

Imagine you own a small gourmet shop that carries special cheeses. You may buy your cheese either domestically—cheddar from New York, Monterey Jack from California—or abroad. If you buy from a domestic firm, it is easy to negotiate the deal and make payment. Because the price of cheese is quoted in dollars and the domestic firm expects payment in dollars, you can pay the same way you pay other bills—by writing a personal check. Only one national currency is involved.

Purchasing cheese from a French cheesemaker is a little more complicated, for two reasons. First, the price of the cheese will be quoted in francs. Second, you will want to pay in dollars, but the French cheesemaker must be paid in francs. Either you must exchange your dollars for francs, or the cheesemaker must convert them for you. At some point, currencies must be exchanged at some recognized exchange rate. **Foreign exchange** is the monetary means or instruments used to make monetary payments and transfers from one currency to another. The funds available as foreign exchange include foreign coin and currency, deposits in foreign banks, and other short-term, liquid financial claims payable in foreign currencies.

International Exchange Rates

Before you buy, you will want to compare the prices of French and domestic cheeses. You must convert the franc price of cheese into its dollar equivalent. To do that, you need to know the international exchange rate between dollars and francs. The **international exchange rate** is the price of one national currency (like the franc) stated in terms of another national currency (like the dollar). In other words, the international exchange rate is the dollar price you must pay for each franc you buy.

Once you know the current exchange rate, conversion of currencies is not difficult. Assume that you want to buy F5,000 (read "5,000 francs") worth of cheese, and that the international exchange rate between dollars and francs is $0.10 (that is, $1 sells for F10). F5,000 at $0.10 apiece will cost you $500. For the rest of this chapter we will assume that the dollar price of the franc is $0.10 to make our arithmetic examples easier to follow.

The international exchange rate determines the dollar price of the foreign goods you want to buy. A different exchange rate would have changed the dollar price of cheese. For instance, suppose the exchange rate rose from $0.10 = F1 to $0.20 = F1. In the jargon of international finance, such a change represents a depreciation (a devaluation involves a depreciation relative to the monetary standard and not necessarily relative to other monies) of the dollar. A **depreciation** of the dollar (or any other national currency) is a reduction in the exchange value or purchasing power, brought about by market forces, in relation to other national currencies. The dollar is now cheaper in terms of francs: It takes fewer francs (F5) to buy a dollar than previously (F10).

The same change represents an appreciation of the franc. An **appreciation** of the dollar (or any other national currency) is an increase in the exchange value or purchasing power, brought about by market forces, in relation to other national currencies. Each franc will now buy a large fraction of a dollar—0.20 as opposed to $0.10. From the perspective of the gourmet shop, the important point is that at the higher exchange rate, the dollar price of the cheese purchase is $1,000 ($0.20 times 5,000). If the exchange rate fell from $0.10 = F1 to $0.05 = F1, the price of the French cheese would decline to $250.

As you can see, your willingness to buy French cheese depends -- much on the franc price of cheese and the exchange rate. If the franc price of cheese increases or decreases, your dollar price increases or decreases.

TABLE 17.7 The Likely Long-Run Effects of Depreciation and Appreciation of the Dollar on U.S. Exports and Imports

	Depreciation Of Dollar	**Appreciation of Dollar**
Price of exports	Decrease	Increase
Total dollar value of exports	Increase	Decrease
Price of imports	Increase	Decrease
Total dollar value of imports	Decrease	Increase

Changes in the dollar price of francs have a similar effect. If the dollar depreciates (that is, if the price of francs in dollars rises), the dollar price of French cheese rises. It is very likely you will be inclined to import less, since at the higher price your customers will buy less. If the dollar appreciates (that is, if the price of francs falls), the dollar price of French cheese falls. Very likely, you will import more because you can lower your own price and sell more. In general, a depreciation of the dollar discourages imports; an appreciation of the dollar encourages imports. The likely long-run results of changes in the international rate of exchange are summarized in Table 17.8 In contrast, in the short run a depreciation can worsen a country's balance of trade according to the J-curve phenomenon because elasticities are smaller. Although the initial impact of depreciation is often an increase in nominal spending on imports because higher prices cause a deterioration in the normal spending on imports, over time depreciation will tend to improve both nominal and real net exports.[8] Thus, although a depreciation in the exchange rate will eventually achieve a balance-of-trade equilibrium as shown in Table 17.8, it may take some time. In general, long-run price elasticities are greater—often considerably greater—than short-run price elasticities. As a rule, economic agents respond reasonably quickly and significantly to changes in economic stimuli.

The Exchange of National Currencies

Assume you have figured the dollar price of cheese using the exchange rate and find it satisfactory. Since your American customers pay for their groceries in dollars, that is the only currency you have to make the payment. Yet cheesemakers in France need francs to pay for their groceries. Therefore the French cheese exporter must ultimately be paid in francs.

How can you make payment in dollars while the French exporter is paid in francs? A bank will exchange your dollars for you. Banks deal in national currencies for the same reason that business people trade in commodities: to make money. An automobile dealer buys cars at a low price with the hope of selling them at a higher price. Banks do the same thing, except that their commodities are national currencies. They buy dollars and pay for them in francs or yen, with the idea of selling them at a profit.

---

[8] Rudiger Dornbush and Paul Krugman, "Flexible Exchange Rates in the Short Run" *Brookings Papers on Economic Activity* (March 1976), pp. 537-575.

If you pay for your French cheese in dollars, you write a check against your checking account and send it to the French firm.[9] The French cheesemaker will accept the check knowing that your dollars can be traded for francs (that is, sold to a French bank) at the current rate of exchange. If the exchange rate is $0.10 = F1, and you have sent the cheesemaker a check for $500, the exporter will receive F5,000 for your check from the French bank. Remember that banks, even foreign ones, have accounts with other banks, just as individuals do. The French bank will deposit your check with its U.S. banker. Your bank balance will fall, and the French bank's balance at the U.S. institution will rise. Then the French bank will sell (or trade) the dollars it has on account for francs.

In the process of buying and selling dollars, the French bank may make a profit. Suppose, for example, that the French bank buys dollars from the French cheesemaker at a rate of $0.10 = F1 (or $1 = F10), paying F500, a net gain of F555.

This hypothetical purchase of French cheese leads to an important observation. Any U.S. import, be it cheese or watches, will increase the dollar holdings of foreign banks. So will American expenditures abroad whether for tours or for foreign stocks and bonds. Americans must have francs for such transactions; therefore, they must offer American dollars in exchange. In most instances, foreign banks end up holding the dollars that Americans have sold.

In the same way, U.S. exports reduce the dollar holdings of foreign banks. Exports are typically paid for out of the dollar accounts of foreign banks. Foreign expenditures on trips to the United States or on the stocks and bonds of U.S. corporations have the same effect. They reduce the dollar holdings of foreign banks and increase the foreign currency holdings of U.S. banks. If American expenditures abroad exceed foreign expenditures here, the dollar holdings of foreign banks will rise—and vice versa.

If American expenditures abroad exceed foreign expenditures here for a long time, foreign banks will eventually accumulate all the dollars they can reasonably expect to use. Foreign banks then have several options. First, they may sell their dollar holdings to other foreign commercial banks to their government—or, more properly, to their government's central bank (for example, the Bank of France).

The market may already be saturated with dollars, however. No one including the central bank, may want to buy dollars at the going price, $0.10 – F1 in our illustration. In that case, foreign banks can induce people to buy dollars by lowering their price. For instance, they can alter the exchange rate from $0.10 = F1 to $0.15 = F1. In so doing they increase the price of francs and decrease (depreciate) the price of dollars.

A depreciation of the U.S. dollar in the exchange rate will have several effects, all tending to reduce the number of dollars coming onto the international money market. As explained earlier, the exchange will make French goods more expensive for Americans to buy. Thus it will tend to reduce U.S. imports, and accordingly the number of dollars that must be exchanged for foreign currencies. Depreciation will also tend to reduce the price of American goods to foreigners. For instance, at an exchange rate of $0.10 = F1, the franc price of a $1 million American computer is F10 million. At an exchange rate of

---

[9] Instruments of exchange other than checks are often used in international transactions. The process, however, is the same.

$0.15 = F1, the franc price of the same computer is F6.66 million—a substantial reduction in price. To buy American goods at the new lower franc price, the French will increase their demand for dollars. Again, the quantity of dollars being offered on the money market will fall, and the growth in foreign dollar holdings will be checked.

Determination of the Exchange Rate

Like the price of anything else, exchange rates are determined by the forces of demand and supply, although government may interfere to alter the rate from what market forces along would have produced. When there is no official or government interference, the rates are free or floating.

When government intervenes, by buying or selling currency in the foreign exchange rates by a central bank or other some official government agency, the exchange rates are fixed or pegged. From 1945 to 1971 exchange rates were basically fixed. Since 1971, however, rates have been set flexibly with some government intervention in a "dirty," or managed, floating exchange rate system, in which the prices of currencies are partly determined by competitive market forces and partly determined by official government intervention.

National currencies have a market value—that is, a price—because individuals, firms, and governments use them to buy foreign goods, services, and securities. There is a market demand for a national currency like the franc. Furthermore, the demand for the franc (or any other currency) slopes downward, like curve *D* in Figure 17.6. To see why, look at the market for francs from the point of view of a U.S. resident. As the dollar price of the franc falls, the price of French goods to Americans also falls. As a result, Americans will want to buy more French goods. They will require a larger quantity of francs to complete their transactions.

The supply of francs coming into the market reflects the French people's demand for American goods, services, and securities. To get American goods, the French need dollars. They must pay for those dollars with francs, and in doing so they supply francs to the international money market. As the dollar price of the franc rises, the price of American goods to the French falls. To buy a larger quantity of American goods at the lower franc price, the French need more dollars; they must offer more francs to get them. Therefore, the quantity of francs supplied on the market rises. Thus the supply curve for francs slopes upward to the right, like curve *S* in Figure 17.6.

The buyers and sellers of francs make up what is loosely called the international money market in francs. Banks are very much involved in such markets. They buy francs from the sellers (suppliers) and sell to the buyers (demanders). As in other markets, the interaction of suppliers and demanders determines the market price. That is, given the supply and demand curves in Figure 17.6, in a competitive market the dollar price of the franc will move toward the equilibrium point at *E* involving the intersection of the supply and demand curves. The equilibrium price, or exchange rate, will be ER_2 , the price at which the quantity of francs supplied exactly equals the quantity of francs demanded.

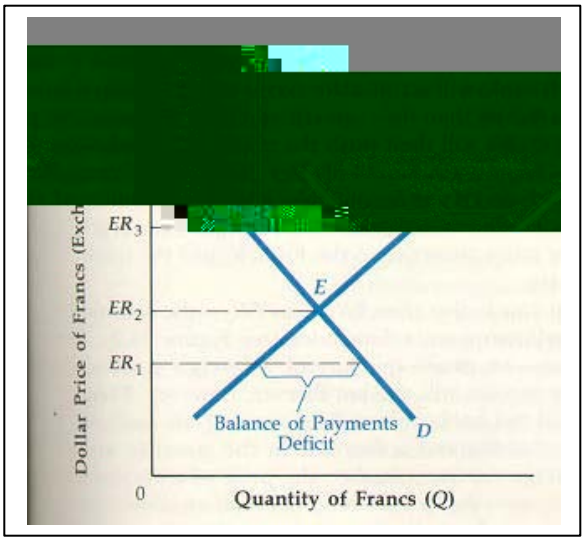


FIGURE 17.6 Supply and Demand for Francs on the International Currency Market

The international exchange rate between the dollar and the franc is determined by the forces of supply and demand with the equilibrium at *E*. If the exchange rate is below equilibrium, say at ER_1, the quantity of francs demanded, shown by the demand curve, will exceed the quantity supplied, shown by the supply curve. Competitive pressure will push the exchange rate up. If the exchange rate is above equilibrium, say at ER_3, the quantity supplied will exceed the quantity demanded, and competitive pressure will push the exchange rate down. Thus the price of a foreign currency is determined in much the say way as the price of any other commodity.

At the market equilibrium point there is no build-up of dollars or francs in the accounts of foreign banks. French and U.S. banks have no reason to modify the exchange rate to encourage or discourage the purchase or sale of either currency. To use a financial expression, the net balance of payments coming into and going out of each nation is zero.

If the exchange rate is below equilibrium level -- say ER_1 -- the quantity of francs demanded will exceed the quantity supplied. An imbalance in the balance of payments will develop. In the jargon of international finance, the United States will develop a balance of payments deficit—a shortfall in the quantity of a foreign currency supplied. (This is a conceptual definition. When it comes to defining the balance of payments deficit in a way that can be measured by the Department of Commerce, economists are in considerable disagreement.)

As in other markets, this imbalance will eventually right itself. Because of the excess demand for francs, French banks will accumulate excess dollar balances. French banks will have more dollars than they can sell and fewer francs than they need. Competitive pressure will then push the exchange rate back up to ER_2. People who cannot buy francs at ER_1 will offer a higher price. As the price of francs rises, French goods will become less attractive to Americans, and the quantity of francs demanded will fall. Conversely, American goods will become more attractive to the French, and the quantity of francs supplied will rise.

Similarly, at an exchange rate higher than ER_2 -- say ER_3 –the quantity of francs supplied will exceed the quantity demanded (see Figure 17.6). A balance of payments surplus—an excess quantity of a foreign currency supplied—will develop. The surplus will not last forever, however. Eventually the exchange rate will fall back toward ER_2, causing an increase in the quantity of francs demanded and a decrease in the quantity supplied. In short, in a free foreign currency market, the price of a currency is determined in the same say the prices of other commodities are determined.

Market Adjustment to Changes in Money Market Conditions

By modifying exchange rates to correct for imbalances in payments, the money market can accommodate vast changes in the economic conditions of nations engaged in trade. A good example is the way the market handles a change in consumption patterns. These changes in consumption, and hence in foreign exchange rates, can be caused by changes in a nation's tastes and preferences, real income, level of prices (including interest rates), costs, and expectations as to future exchange rates. If all countries' exchange rates move with the relative rates of inflation, only real (terms of trade) changes would affect the relative prices of home country to foreign-country goods. However, while floating exchange rates tend to eliminate automatically any balance-of-payment problems, they may diminish the volume of trade because of the uncertainty and instability of the terms of trade. In fact, since flexible exchange rates were reintroduced in 1971 the volume of world trade has actually grown despite considerable volatility and turbulence.

The two major advantages of a floating system are that exchange rates are automatically determined exclusively by free market forces, without government intervention, controls, or regulations. Moreover, external adjustment, under favorable conditions, is attained without requiring major domestic or internal price, income, or employment changes. Its two major disadvantages are: (1) uncertainty and instability in the form of frequent and large fluctuations discourages international trade, transactions, and investment; and (2) there is the possibility of exchange rate fluctuations leading to cumulative disequilibrium rather than stable equilibrium.

Suppose American preferences for French goods—say, wines and perfumes—increase for some reason. The demand for francs will rise because Americans will need more francs to buy the additional French goods they desire. If, as in Figure 17.7, the U.S. demand for francs shifts from D_1 to D_2, the quantity of francs demanded at the old equilibrium exchange rate of ER_1 will exceed the quantity supplied. Those who cannot buy more francs at ER_1 will offer to pay a higher price. The exchange rate will rise toward the new equilibrium level of ER_1 as the equilibrium point shifts from E_1 to E_2. As the dollar depreciates in value, the imbalance in payments is eliminated.

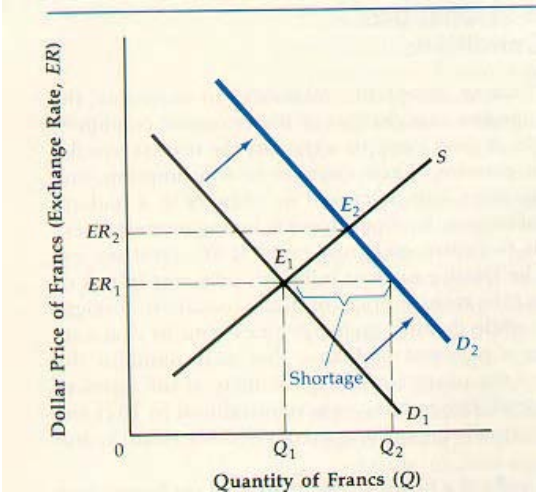


FIGURE 17.7 Effect of an Increase in Demand for Francs

An increase in the demand for francs will shift the demand curve from D_1 to D_2, pushing the equilibrium from E_1 to E_2. At the initial equilibrium exchange rate ER_1, a shortage will develop. Competition among buyers will push the exchange rate up to the new equilibrium level ER_2.

Now suppose Americans' real incomes rise. Assuming that the consumption of goods and services goes up with real income—we called these "normal" goods and services earlier in the book – Americans will be likely to demand more foreign imports, both directly and in the form of domestic goods that incorporate foreign parts or materials. Either way, an increase in real incomes leads to an increase in the demand for foreign currencies. Again the demand for francs will rise, as in Figure 17.7. The exchange rate will rise with it to bring the quantity supplied into line with the quantity demanded.

A change in the rate of inflation can have a similar effect on the exchange rate. If the inflation rates are about the same in two nations that trade with each other, the exchange rate between their currencies will remain stable, *ceteris paribus*, according to the **purchasing power parity theory**. Because the relative prices of goods in the two nations stay the same, people will have no incentive to switch from domestic to imported goods, or vice versa. If one nation's inflation rate exceeds another's, however, the relative prices of foreign and domestic goods change. If prices increase faster in the United States, for example, Americans will want to buy more foreign goods and fewer domestic goods. Foreigners, on the other hand, will have an incentive to buy more goods from their own countries, where prices are not rising as fast as in the United States. In sum, a higher U.S. inflation rate spells a rise in the demand for foreign currencies, a fallen in their supply, and a depreciation of the dollar. Similar flows occur when there are interest rate differentials between nations.

Figure 17.8 illustrates the process for prices in general. As U.S. demand for foreign goods rises, the demand curve for francs shifts outward from D_1 to D_2, shifting the equilibrium from E_1 to E_2. As foreign demand for U.S. products falls, the supply curve for francs shifts to the left, from S_1 to S_2. At the initial equilibrium exchange rate of ER_1, a shortage of francs will develop. The exchange rate will rise to ER_2, eliminating the shortage and reestablishing balance in the money market. At the higher rate, Americans must pay a higher dollar price for foreign goods. The rise in the exchange rate has evened out the difference in the two nations' inflation rates.

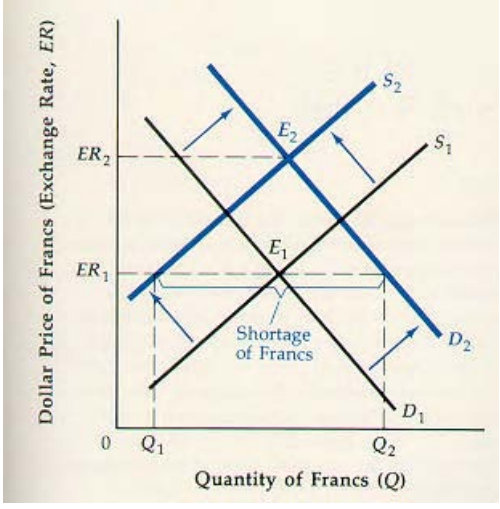


FIGURE 17.8 Effect of an Increase in Inflation on the Supply and Demand for Francs

If the rate of inflation is higher in the United States than in France, the demand for francs will rise from D_1 to D_2, while the supply of francs will contract from S_1 to S_2. The dollar price of francs will rise from ER_1, to ER_2, as the equilibrium shifts from E_1, to E_2.

In the short-run, supply and demand are most influenced by anticipations as to the direction in which an exchange rate is likely to move. For example, if the franc is expected to increase in value, people who have payments to make in that currency will tend to buy the currency and make payments sooner. Economic and political news—such as an unanticipated change in monetary policy—has an almost immediate impact.

Control of the Exchange Rate: The Fixed or Pegged Rate System

So far our analysis of the international money market has assumed a floating, or flexible, system of exchange in which exchange rates are determined by private demand and supply forces in the market. A **floating, flexible, or freely fluctuating exchange rate system** is one in which the prices of currencies are determined by competitive market forces. Until 1971, however, international exchange rates were controlled by governments. Rates were not permitted to move in response to changes in supply and demand. Because rates were fixed for long periods of time by government decree, this system is generally referred to as a fixed exchange rate system. A **fixed or pegged exchange rate system** is one in which the prices of currencies are established and maintained by government intervention. Although the fixed-rate system is no longer in use among major nations, it merits some discussion because of its historical importance and because of periodic high-level discussions—especially in the late 1980s—about returning to it.

To understand that a properly working fixed exchange rate system can be better than a floating-rate system, consider the problems that would arise if *each state* in the United States had its own currency. The exchange rate would vary among all the states. The resulting risks and inconveniences would severely hamper interstate trade. For instance, a worker in New York City who commutes from New Canaan, Connecticut, would have to face fluctuating exchange rates on a daily basis when riding subways, buying gas, eating lunch, whatever.

The fixed exchange rate has one advantage over the floating rate: it is stable. Because even a small change in the exchange rate can cause significant losses to people who have already concluded business deals, a flexible exchange rate can increase the risks involved in international trade. For example, suppose you agree to purchase cheese at an exchange rate of $0.10 = F1. You promise to pay the exporter $00, and the French cheesemaker expects to receive F5,000. By the time you send the check, however, the rate has moved to $0.11 = F1. The exporter will now receive only F4,545 ($500 ÷ 0.11). She loses F455.

If the exchange rate moves in the opposite direction, of course, the exporter will gain. In addition, the French cheesemaker can hedge against short-term losses by agreeing, at the time she closes the deal, to sell the proceeds at a given exchange rate, perhaps a fraction of a cent less than the current rate of $.10 = F1. In long-term deals, however, traders inevitably risk losing money because of changes in exchange rates.

They incur a risk cost that is translated into higher prices. Under a fixed-rate system, exchange rates move only periodically. The risk cost is reduced, and the prices of foreign goods can be lower.

Like any other form of price control, however, control of foreign exchange rates creates its own problems. If the exchange rate is fixed—at ER_1 in Figure 17.8, for example—and the supply and demand curves remain stable, there is no problem. There is no need for government to fix the rate either, however, It will remain ER_1 as long as the supply and demand curves for currency stay put.

Problems can develop when market conditions change but the exchange rate is fixed. If the demand for francs increases from D_1 to D_2 in Figure 17.8, a shortage of francs will develop on the international money market. Those who want francs at the fixed price will be unable to get all they want. The government may have to ration the available francs and police the market against black marketeering. If black markets are not controlled, the price of currency will rise—illegally perhaps, but it will rise nonetheless. In the end, the exchange rate will not really be controlled.

Perhaps the chief disadvantage of a fixed rate system is that the level of internal prices and costs in each nation is affected by external economic and monetary developments over which a nation has little or no control. Nations must play according to the rules of the game and submit their internal economy to the dictates of external equilibrium.

Concluding Comments

The schedule of tariffs applied to goods coming into the United States is now larger than the Los Angeles telephone directory. Surely all those tariffs were not imposed in pursuit of the national interest, as in the maintenance of a strong defense industry. Most probably reflect the political influence of special-interest groups. Yet on balance, the overall tariffs are low, but they mask very high tariffs and even quotas on certain commodities—such as certain agricultural products, tobacco, motorcycles, and cooking utensils.

The case against such special-interest tariffs was wittily stated by the nineteenth-century French economist Frederic Bastiat. Pretending to represent the candle manufacturers of his day, he wrote to the French Chamber of Deputies in 1845:

> Gentlemen:
>
> . . .We are subjected to the intolerable competition of a foreign rival, who enjoys, it would seem such superior facilities for the production of light that he is enabled to *inundate* our *national market*, at so exceedingly reduced price, that, the moment he makes his appearance, he draws off all customs for us; and thus an important branch of French industry . . . is suddenly reduced to a state of complete stagnation. This rival is no other than the sun.
>
> Our petition is, that it would please your honorable body to pass a law whereby shall be directed the shutting up of all windows, doors, skylights, shutters,

> curtains, in a word, all opening, holes, chinks, and fissures through which light of the sun . . . penetrates into our dwellings.[10]

Bastiat suggests that passage of his proposed law would be consistent with the chamber's attempts to check the importation of "coal, iron, cheese, and goods of foreign manufacture, merely because and even in proportion as their price approaches zero."

Clearly, tariffs force consumers to pay more for domestic goods. In that extent they reduce aggregate real income. Unfortunately because they benefit special-interest groups—tariffs, like other taxes, are probably inevitable.

Review Questions

1. Using supply and demand curves, show how a U.S. tariff on a foreign-made good will affect the price and quantity sold in the country of origin.
2. How will an import quota on sugar affect the price of sugar produced and sold domestically? Sugar produced domestically and sold abroad?
3. If a tariff is imposed on imported autos and the domestic demand for autos rises, what will happen to auto imports? If a quota is imposed on imported autos and the demand for autos increases, what will happen to auto imports?
4. Given the following production capabilities for cheese and bread, which nation will export cheese to the other? What might be a mutually beneficial exchange rate for cheese and bread?

	Cheese		**Bread**
France	40 units	or	60 units
Italy	10 units	or	5 units

5. "Tariffs on imported textiles increase the employment opportunities and incomes of domestic textiles workers. They therefore increase aggregate employment and income." Evaluate this statement.
6. Since the balance of payments must always balance, how can a disequilibrium situation occur?
7. How much would a business spend to get a tariff? What economic considerations will have an impact on the amount?

---

[10] Frederic Bastiat, "A Petition," *Economic Sophisms* (Irvington-on-Hudson, N.Y., Foundation for Economic Education, 1964; originally published 1945), purchasing power. 56-60.

关于本文档

About This Document

1. 本文档由七彩英语网站(www.qcenglish.com)整理发布，文档内容均来自互联网，文档中的观点并不代表七彩英语网站立场
2. 本文档版权归原版权所有人所有，任何涉及商业盈利目的均不得使用，否则产生的一切后果将由您自己承担
3. 整理发布本文档是出于让更多的人有机会接触原版书籍的目的，如果喜欢，请购买正版
4. 欢迎广大英语爱好者分享传播本文档，让更多的英语爱好者可以阅读原版书籍
5. 更多精彩原版书籍请登录七彩英语网站
6. 如果大家对七彩英语网站发布的电子书有任何意见和建议，请不吝赐教

1. This document is collected from the Internet and distributed by QCENGLISH. It doesn't mean that QCENGLISH agrees to all the views in this document.
2. Without the permission from the copyright owner business use is forbidden, or you should take all the responsibilities.
3. We collect and distribute this document on the purpose of sharing good English books with more English fans. If you like this document please buy a genuine one.
4. It is recommended that English fans share this document with anyone who wants to improve himself(herself) through reading English books.
5. More books can be downloaded from QCENGLISH.
6. Please don't hesitate to contact us if you have any idea about the document we distributed.