ECONOMICS OF REGULATION AND ANTITRUST

Third Edition

W. Kip Viscusi
John M. Vernon
Joseph E. Harrington, Jr.

The MIT Press
Cambridge, Massachusetts
London, England

f. It shall be unlawful for any person engaged in commerce, in the course of such commerce, knowingly to induce or receive a discrimination in price which is prohibited by this section.

3. It shall be unlawful for any person engaged in commerce, in the course of such commerce, to lease or make a sale or contract for sale of goods, wares, merchandise, machinery, supplies, or other commodities, whether patented or unpatented, for use, consumption, or resale within the United States or any Territory thereof or the District of Columbia or any insular possession or other place under the jurisdiction of the United States, or fix a price charged therefor, or discount from, or rebate upon, such price, on the condition, agreement, or understanding that the lessee or purchaser thereof shall not use or deal in the goods, wares, merchandise, machinery, supplies, or other commodities of a competitor or competitors of the lessor or seller, where the effect of such lease, sale, or contract for sale or such condition, agreement, or understanding may be to substantially lessen competition or tend to create a monopoly in any line of commerce.

7. No corporation engaged in commerce shall acquire, directly or indirectly, the whole or any part of the stock or other share capital and no corporation subject to the jurisdiction of the Federal Trade Commission shall acquire the whole or any part of the assets of another corporation engaged also in commerce, where in any line of commerce in any section of the country, the effect of such acquisition may be substantially to lessen competition, or to tend to create a monopoly. This section shall not apply to corporations purchasing such stock solely for investment and not using the same by voting or otherwise to bring about, or in attempting to bring about, the substantial lessening of competition. Nor shall anything contained in this section prevent a corporation engaged in commerce from causing the formation of subsidiary corporations for the actual carrying on of their immediate lawful business, or the natural and legitimate branches or extensions thereof, or from owning and holding all or a part of the stock of such subsidiary corporations, when the effect of such formation is not to substantially lessen competition.

Federal Trade Commission Act

5. a. (1) Unfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are declared unlawful.

(2) The Commission is empowered and directed to prevent persons, partnerships, or corporations, except banks, common carriers subject to the Acts to regulate commerce, air carriers and foreign air carriers subject to the Federal Aviation Act of 1958, and persons, partnerships, or corporations insofar as they are subject to the Packers and Stockyards Act, 1921, as amended, except as provided in section 406 (b) of said Act, from using unfair methods of competition in or affecting commerce and unfair or deceptive acts or practices in or affecting commerce.

4 Efficiency and Technical Progress

As indicated in the preceding chapter, *economic performance* is the term used to measure how well industries accomplish their economic tasks in society's interests. Clearly, to evaluate antitrust laws it is essential to have some well-defined objective. In order to evaluate a law that prohibits mergers between two rivals, it is important to have a conceptual tool that identifies the costs and benefits to society of that law.

The two dimensions of economic performance to be discussed here were referred to in the last chapter as *efficiency* and *technical progress.* In a sense, more descriptive terms would be static and dynamic efficiency—but we use the traditional terms in order to be consistent with the economics literature. The main distinction is that in discussing efficiency it will be assumed that the technology is given, and in discussing technical progress the assumption is that resources are being allocated to developing new technologies (for producing old products more cheaply and for producing completely new products).

Economic Efficiency

We begin by considering the theoretical world of perfect competition. Every microeconomics text devotes much attention to the perfectly competitive model. The key assumptions are these:

1. Consumers are perfectly informed about all goods, all of which are private goods.
2. Producers have production functions that rule out increasing returns to scale and technological change.
3. Consumers maximize their preferences given budget constraints; producers maximize profits given their production functions.
4. All agents are price takers, and externalities among agents are ruled out.
5. A competitive equilibrium, that is, a set of prices such that all markets clear, is then determined.

An important welfare theorem that follows from the preceding assumptions is that the competitive equilibrium is *Pareto optimal.* In short, the equilibrium cannot be replaced by another one that would increase the welfare of some consumers without harming others. An important property of the equilibrium is that *price equals marginal cost* in all markets.

Note that the ideal competitive world that we have described would have no need for government intervention in the marketplace, except for policies affecting income distribution. This book ignores problems of income distribution—leaving those problems to the field of public finance (which studies taxation and transfer payments).

Many of the listed assumptions will be relaxed and discussed in detail throughout this book. Of course, the key assumption to be discussed in this part of the book is the *price-taking* assumption. That is, antitrust economics is concerned with the causes and consequences of firms' abilities to set price above marginal cost.

Once we begin to relax these assumptions, it becomes clear that we need to develop partial equilibrium tools. That is to say, it becomes incredibly complex to deal with a general equilibrium model in which some markets are monopolies, externalities exist, imperfect information about product quality obtains, and so on.[1] Hence we now turn to welfare economics concepts in the context of a single market, effectively ignoring the interactions with all other markets.

Partial Equilibrium Welfare Tools

The competitive model described by the list was said to satisfy the condition of *Pareto optimality.* This is also referred to as *Pareto efficiency* or simply *economic efficiency.* One tool for evaluating the effect of a policy change (say, breaking up a monopoly) is the Pareto criterion. That is, if everyone is made better off by the change (or no one is made worse off, and at least one person is made better off), then the Pareto criterion would say that the change is "good." It is hard to argue with this criterion for evaluating public policies. The problem is that one is unlikely to find many "good" real-world policies. In most cases in the real world, at least some people will be harmed.

A generally accepted alternative standard in applied microeconomics is the *compensation principle,* which is equivalent to choosing policies that yield the highest total *economic surplus.* The basic idea is that if the "winners" from any policy change can, in principle, compensate the "losers" so that everyone is better off, then it is a "good" change. Note that actual compensation of the losers is not required. If it were required, of course, it would satisfy the Pareto criterion.

To illustrate, consider Figure 4.1. The figure shows the market demand and supply curves for videocassette recorders (VCRs). Recall first a few facts about these two curves. The competitive industry's supply curve is found by horizontal aggregation of the supply curves of individual firms. The individual firms' supply curves are their marginal cost curves; hence we can think of the supply curve in Figure 4.1 as the industry's marginal cost curve.

Another useful point is to recognize that the area under the marginal cost curve represents the sum of the incremental costs for all units of output and, as a result, equals the total cost. Hence the total cost of producing Q^* VCRs is the area $0Q^*DC$ (this is exclusive of any fixed costs).

Under certain assumptions, the demand curve can be viewed as a schedule of the marginal *willingness-to-pay* by VCR customers.[2] For example, at the competitive equilibrium (price P^*, output Q^*) the marginal willingness-to-pay P^* exactly equals marginal cost at the output Q^*. Because the area under this schedule of marginal willingness-to-pay is total willingness-to-pay, consumers are willing to pay $0Q^*DA$ for output Q^*. The difference between total willingness-to-pay and total cost is therefore the area ACD and is referred to as the *total surplus* generated in the VCR market. Finally, it is common to divide total surplus into *consumer surplus* of AP^*D and *producer surplus* of P^*CD.

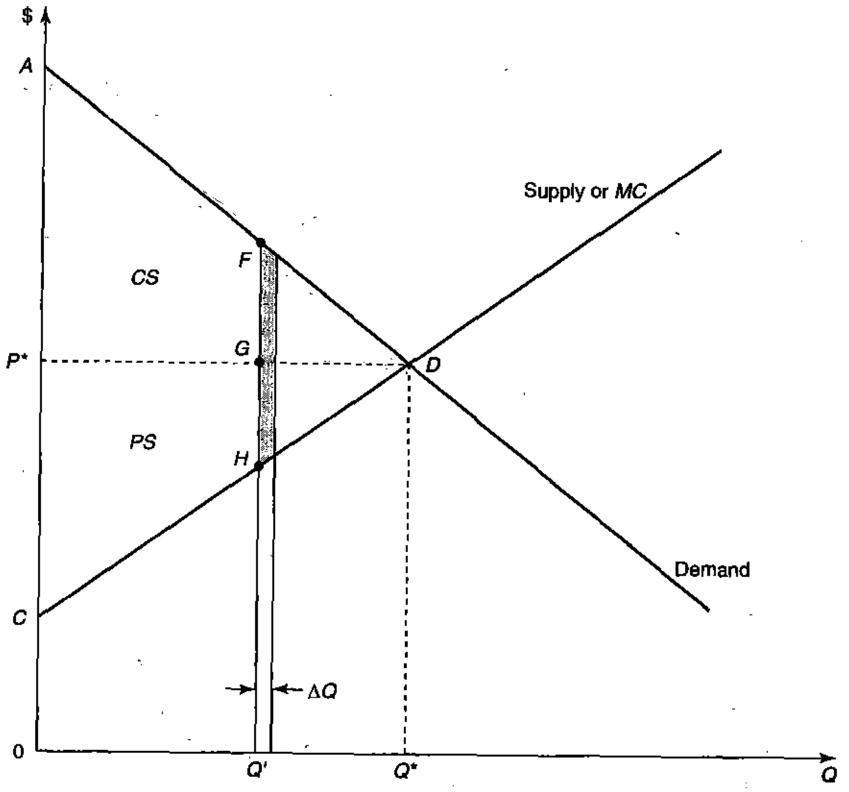

Figure 4.1
Demand and Supply Curves in Determination of Economic Surplus

[1] See R. G. Lipsey and K. Lancaster, "The General Theory of Second Best," *Review of Economic Studies,* 1956, for an analysis.

[2] This interpretation is most easily understood if the demand curve is assumed to be made up of many heterogeneous consumers with demands for at most one VCR. Hence the individual with the highest valuation (or willingness-to-pay) for a VCR is represented by the vertical intercept of the demand curve, $0A$. The next highest valuation (for the second VCR) is slightly less than $0A$, and so forth. The person who actually has the marginal willingness-to-pay P^* is the person who obtains a zero (individual) consumer surplus—all others have positive surpluses. For example, the person with marginal willingness-to-pay of $Q'F$ has to pay P^* and has a surplus of FG. The key assumption necessary to make this interpretation generally valid is that the income effect for the good is "small." See R. Willig, "Consumer's Surplus without Apology," *American Economic Review,* September 1976, for support for this interpretation.

Consumer surplus is defined as the total willingness-to-pay $0Q^*DA$ less what the consumers must actually pay. Because consumers must pay the rectangle defined by price P^* and the output Q^* (that is, area $0Q^*DP^*$), the area AP^*D in Figure 4.1 is the consumer surplus. Producer surplus, defined in an analogous manner, is equal to the profit of the firms in the industry. Because firms receive revenues of price P^* times output Q^* (that is, area $0Q^*DP^*$) and they incur costs equal to the area under the marginal cost curve, $0Q^*DC$, they earn a producer surplus of the difference, P^*CD.

Notice that maximizing total surplus is equivalent to maximizing the sum of consumer and producer surplus. We next show that maximizing total surplus is equivalent to selecting the output level at which price equals marginal cost. In Figure 4.1, assume that output Q' is being produced and sold at price $Q'F$. Clearly, at the output Q', the marginal willingness-to-pay $Q'F$ exceeds the marginal cost $Q'H$. Hence a small increase in output of ΔQ would increase surplus by the area of the slender shaded region (approximately FH height by ΔQ width). Output increases would continue to increase surplus up to output Q^*. Hence, maximizing surplus implies that output should be increased from Q' to Q^*, adding an increment to total surplus of area FHD. Of course, by an analogous argument, we can show that output increases beyond Q^* would reduce surplus, since marginal cost exceeds marginal willingness-to-pay. In short, equating price and marginal cost at output Q^* maximizes total surplus.

It is useful to provide another interpretation for the area FHD in Figure 4.1. Recall that this area represents potential increases in total surplus if for some reason output is held at Q'. For illustrative purposes, assume that a cartel has agreed to restrict output to Q', charging price $Q'F$. This results in a so-called *deadweight loss* of surplus equal to area FHD. This is often referred to as the *social cost of monopoly*, or simply an *efficiency loss*. In other words, without the cartel, competition would cause price to equal marginal cost, yielding the higher total surplus of ACD as compared to the surplus under the cartel case of $ACHF$. As before, it is sometimes said that there is a deadweight loss in consumer surplus of the triangle FGD and a deadweight loss of producer surplus of the triangle GHD.

Now, consider the point made earlier about the compensation principle and the argument that if the winners can compensate the losers the policy change is a good one. Using a simple monopoly-versus-competition example, we will show that additional insights can be obtained by considering consumers and producers separately.

Monopoly-versus-Competition Example

In Figure 4.2 we show a monopoly equilibrium with price P_m and quantity Q_m. For simplicity, we assume that average cost AC is constant and therefore equal to marginal cost MC. Hence the monopolist chooses output Q_m where marginal revenue MR equals marginal cost. Profit, or producer surplus, equals price minus average cost multiplied by quantity, or area P_mP_cCB. Consumer surplus equals the triangle AP_mB.

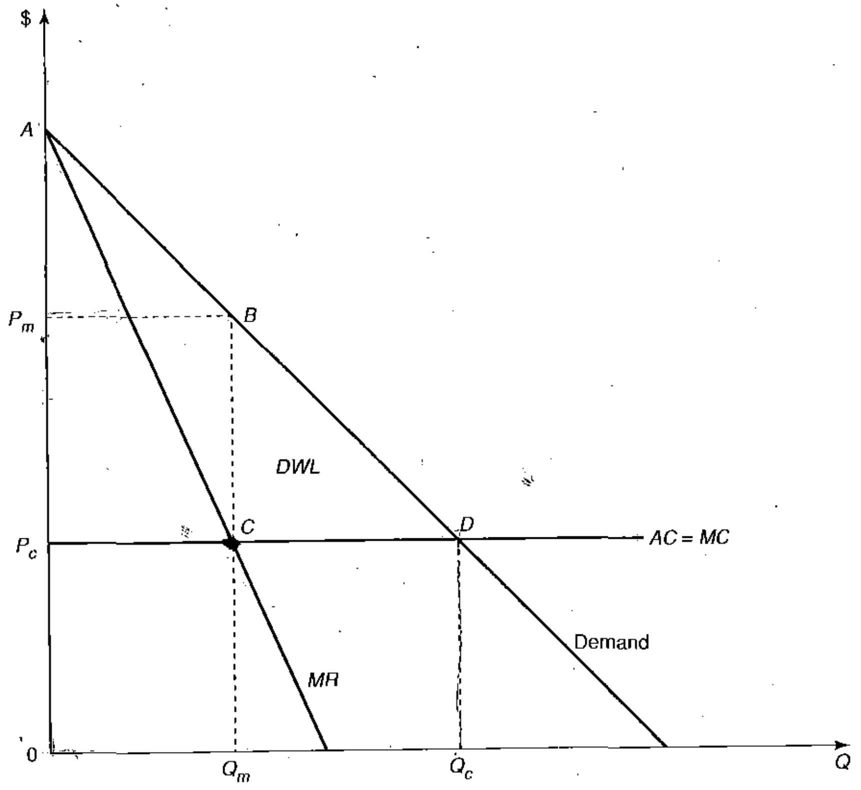

Figure 4.2
Monopoly versus Competition

Next, consider a policy to break up the monopoly and replace it with a competitive industry. Let us assume no change in costs, so that the competitive industry supply is the horizontal line at the level of MC. (This assumption may not be satisfied in practice, inasmuch as one reason for the existence of a monopoly may be some technological superiority that achieves lower costs of production.) Hence the new equilibrium is price P_c and output Q_c. Consumer surplus increases to the triangular area AP_cD, and producer surplus disappears.

In effect, the elimination of monopoly has led to a net gain in total surplus of triangle BCD. This triangle, the deadweight loss caused by the monopoly, is labeled as DWL in Figure 4.2.

To reinforce the points we have made, we can use specific numerical demand and cost functions. In particular, assume

$$Q = 100 - P \qquad \text{Demand}$$

$$MC = AC = 20 \qquad \text{Marginal and average cost.}$$

The monopoly price is therefore $P_m = \$60$, $Q_m = 40$, and the competitive equilibrium is $P_c = \$20$, $Q_c = 80$.[3]

Monopoly

$$\text{Total surplus} = AP_cCB = \$2{,}400$$

$$\text{Consumer surplus} = AP_mB = \$800$$

$$\text{Producer surplus} = P_mP_cCB = \$1{,}600$$

Competition

$$\text{Total surplus} = AP_cD = \$3{,}200$$

$$\text{Consumer surplus} = AP_cD = \$3{,}200$$

$$\text{Producer surplus} = 0$$

The procompetition policy leads to an increase in total surplus from $2,400 to $3,200. On this basis, it should be carried out. Notice, however, at the disaggregated level, producer surplus falls from $1,600 to zero. The owners of the monopoly are therefore harmed. Consumers gain enough to compensate the monopoly owners and still be better off. That is, consumers gain by $\$3{,}200 - \$800 = \$2{,}400$. In principle, consumers could compensate the monopoly owners with $1,600 to offset their loss, and still have a net gain of $\$2{,}400 - \$1{,}600 = \$800$. Of course, as discussed earlier, under the compensation principle the compensation need not be carried out. One can justify this outcome by noting that if the government is worried about the income level of the monopoly owners, it can handle this concern directly through the tax system.

Oil Industry Application

An interesting application of this type of analysis of the oil industry was performed by Arrow and Kalt[4] in 1979. They evaluated the benefits and costs of removing oil price controls in the United States. While the controls will be examined in detail in Chapter 18, it is instructive to present their main findings here to illustrate efficiency losses and gains as compared with simple transfers of surplus from one group to another.

In the 1970s the federal government, concerned with inflation, held oil prices in the United States below what prices would have been in the absence of the controls. This resulted in efficiency losses, according to Arrow and Kalt, of approximately $2.5 billion per year. (A detailed analysis of these losses is provided in Chapter 18.)

Our preceding analysis, shared by most economists, is that this is as far as economists can legitimately go in evaluating public policies. It then becomes a political decision as to whether the transfers among groups are viewed as supporting or offsetting the efficiency analysis. For example, in the hypothetical monopoly example, the transfer of surplus is from the monopoly owners to consumers, and this is presumably in the politically "correct" direction. That is, if one believes that consumers generally have lower incomes than monopoly owners, and that a more equal income distribution is good, breaking up the monopoly both eliminates efficiency losses and has politically correct distribution effects.

Arrow and Kalt took a further step by trying to evaluate the distribution effect of decontrolling oil prices. Roughly, the decontrol of oil prices would mean higher prices for consumers and higher profits for producers—a politically bad transfer. They were concerned with trying to compare the gain in efficiency with the loss in equity.

The transfer from consumers to producers was estimated to be about $2.8 billion. Arrow and Kalt then proposed, with numerous qualifications, that a dollar transfer from consumers to producers would lose about half its value. The resulting "equity cost" as they termed it would then be half of the $2.8 billion transfer, or $1.4 billion. Hence the efficiency gain of $2.5 billion[5] exceeded the equity cost of $1.4 billion, and they therefore recommended that oil price decontrol was in the public interest.

The key to Arrow and Kalt's analysis is their willingness to assign an "equity cost" of 50 cents per dollar transferred from consumers to producers. As noted earlier, the standard view of economists is that assigning an equity cost of this sort is arbitrary. Economic analysts currently have no empirical basis for assigning any specific value to these equity costs. Nevertheless, it is certainly true that the political process gives great weight to equity issues, and it is helpful for economists to at least set out the magnitude involved.

Some Complications

Economies of scale were implicitly assumed to be relatively small in the monopoly-versus-competition example. That is, we ignored the problem that arises when the representative firm's long-run average cost curve reaches its minimum at an output level that is large relative to the market demand. In other words, in our monopoly example, we assumed that the single firm could be replaced with a large number of firms with no effect on costs.

[3] The monopolist sets marginal revenue MR equal to MC. MR is $100 - 2Q$ and MC is 20. Equating and solving for Q gives $Q = 40$. The competitive equilibrium is found by setting $P = MC$. So $100 - Q = 20$ gives $Q = 80$. In each case substitute the equilibrium value of Q into the demand function to obtain the value of P.

[4] K. J. Arrow and J. P. Kalt, "Decontrolling Oil Prices," *Regulation*, September/October 1979.

[5] Actually, Arrow and Kalt noted that the $2.5 billion efficiency gain from decontrol should be reduced to $1.9 billion to reflect the fact that the efficiency gains would accrue primarily to producers. Thus the final comparison was a $1.9 billion gain and a $1.4 billion loss in favor of decontrol.

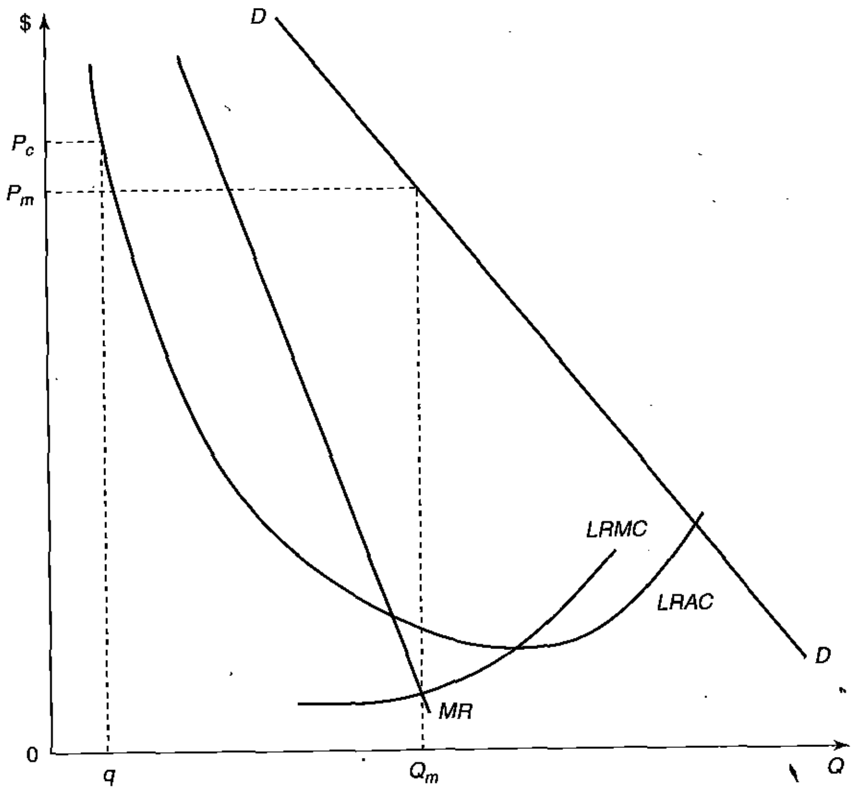

Figure 4.3
Economies of Scale and Natural Monopoly

To take an extreme case, consider Figure 4.3. Economies of scale are such that the long-run average cost curve $LRAC$ reaches its minimum at an output level that is very large relative to market demand. Situations of this kind are referred to as *natural monopolies,* to reflect that production can be most cheaply carried out by a single firm. The profit-maximizing monopolist would set price equal to P_m and output Q_m.

Suppose that it were known that in order to have a sufficient number of firms in the industry for competition to obtain, each firm would be able to produce an output of only q. As Figure 4.3 shows, the average cost of output q would be quite high and would result in a competitive price of P_c, *which exceeds the monopoly price.*

Clearly, economies of scale can make monopoly the preferred market organization. Public utilities to provide electric power or sewage treatment are notable examples. In extreme cases of the type depicted in Figure 4.3, the policy problem becomes one of regulating the natural monopolist. The approach usually followed in public utility regulation is to force the monopolist to price so as to earn a "fair" rate of return on its investment. An alternative, although not often followed in the United States, is to create a public enterprise, owned and operated by the government. These topics will be discussed in detail in Part II.

More relevant to antitrust policy is the intermediate case, where economies of scale are more moderate relative to market demand. For example, it may be imagined that the size of the automobile market is only large enough to support three or four firms, each producing at the minimum point on its long-run average cost curve. This situation would give rise to an industry of three or four firms, or an oligopoly. The key factor differentiating oligopoly from perfect competition and monopoly is that the small number of firms creates a high degree of interdependence. Each firm must consider how its rivals will respond to its own decisions.

Oligopoly theory does not yield any definite predictions analogous to the *price = marginal cost* prediction of perfect competition, or the *price greater than marginal cost* prediction of monopoly. Most theories of oligopoly imply that price will exceed marginal cost, but by less than under monopoly.

Yet oligopoly is quantitatively very significant in most industrial economies, and it is therefore an important topic for study. It should be stressed, in addition, that the prevalence of oligopoly does not necessarily imply that large-scale economies are the cause. In fact, whether or not economies of scale explain the existence of particular oligopolies is a key public policy concern. We will return to oligopoly theory in Chapter 5.

A second complication is the existence of *product differentiation.* Product differentiation refers to the situation in which some differences in the products of rival sellers are perceived by the buyers. The differences may be real differences, such as the differences in size, styling, horsepower, reliability, and so on, between Fords and Chevrolets—or they may be primarily the result of image differences conveyed through advertising. The main requirement is that consumers regard the differentiation sufficiently important that they willingly pay a somewhat higher price for their preferred brand.

E. H. Chamberlin[6] constructed the theory of monopolistic competition in which many competitors produce differentiated products. All firms that produce products that are reasonably close substitutes are members of the *product group.* Given these assumptions and the assumption of free entry, the long-run equilibrium of a monopolistic competitor is given by the tangency of the firm's demand curve with its average cost curve. This is shown in Figure 4.4.

The monopolistic competitor earns zero profits in long-run equilibrium. This is a consequence of the assumption of free entry; the existence of a positive profit will attract entry until a firm's own demand is reduced sufficiently to make profits zero. The product differentiation assumption gives the firm's demand curve its slightly negative slope; that is, the firm can increase its price without losing all its sales to a competitor.

[6] E. H. Chamberlin, *The Theory of Monopolistic Competition* (Cambridge, Mass.: Harvard University Press, 1933).

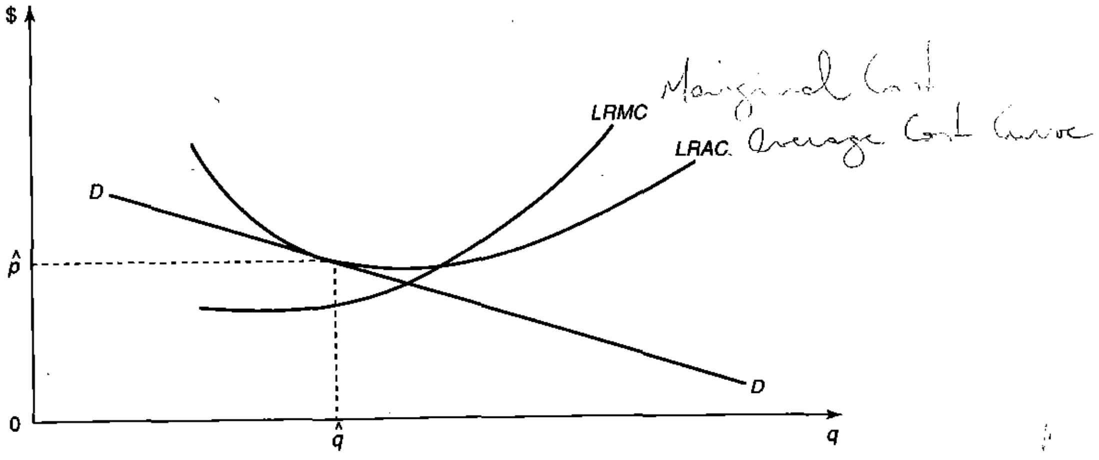

Figure 4.4
Equilibrium of Monopolist Competitor

The relevant point here is that price exceeds marginal cost—the signal that there is a misallocation of resources. But consider Chamberlin's argument:

> The fact that equilibrium of the firm when products are heterogeneous normally takes place under conditions of falling average costs of production has generally been regarded as a departure from ideal conditions. . . . However, if heterogeneity is part of the welfare ideal, there is no prima facie case for doing anything at all. It is true that the same total resources may be made to yield more units of product by being concentrated on fewer firms. . . . But unless it can be shown that the loss of satisfaction from a more standardized product is less than the gain through producing more units, there is no "waste" at all, even though every firm is producing to the left of its minimum point.[7]

The key issue is the optimal amount of product variety, and this is a difficult theoretical problem. A large literature on this subject has developed since Chamberlin's observation.[8] In Chapter 6 we present a simple model that illustrates the trade-offs involved.

X-Inefficiency

Other types of inefficiency may be important in monopoly. First, we consider X-inefficiency, so named by Leibenstein in his well-known 1956 article on the subject.[9] Thus far, we have assumed that both monopolists and perfect competitors combine their factors of production efficiently, thereby minimizing cost for each level of output. However, it can be argued that the pressures of competition force perfect competitors to be cost minimizers, whereas the freedom from competition makes it possible for the monopolist to be inefficient, or X-inefficient. That is, the monopolist may operate at a point *above* its theoretical cost curve.

Of course, X-inefficiency is inconsistent with the assumption that monopolists maximize profits. However, some economists have argued that the separation of ownership from control in large firms with market power permits the managers to substitute their own objectives for the profit objectives of the owners. Therefore, in such cases, X-inefficiency may arise.

Monopoly-Induced Waste

A third and final source of inefficiency created by monopoly is competition among agents to become a monopolist. Consider the example of a government-mandated monopoly in the form of a franchise. If Figure 4.2 depicts the relevant demand and cost curves, then the franchise owner will earn profits equal to $P_m P_c C B$. Knowing that the firm that receives this franchise will earn rents of $P_m P_c C B$, firms will invest resources in lobbying the legislature or the regulatory agency in order to become the recipient of this franchise. This competition to earn monopoly profits uses up real resources in the form of labor by lobbyists and lawyers. These wasted resources represent a cost to society, just as do the traditional deadweight loss and any X-inefficiencies. Competition among firms for rents is appropriately referred to as *rent-seeking* behavior.[10]

How large is the welfare loss from rent-seeking behavior? We know that it cannot exceed the amount of monopoly profits ($P_m P_c C B$ in Figure 4.2). No firm would find it optimal to spend in excess of that amount in order to become a monopolist. In some simple models it has been shown that if rent-seeking is perfectly competitive (that is, there are many identical firms), then all rents will be competed away.[11] In that case, the total welfare loss from monopoly is $P_m P_c D B$. More generally, $P_m P_c D B$ represents an upper bound on the welfare loss from monopoly (excluding any X-inefficiencies) while BCD is a lower bound.

There are a number of ways in which rent-seeking behavior may arise. As just mentioned, competition for rents could take the form of firms lobbying legislators in order to get favorable legislation passed, for example, entry regulation and import quotas. When these lobbying activities use up real resources, they represent a welfare loss associated with monopoly. Alternatively, if favorable government actions are achieved by bribing legislators or regulators, then this is not a welfare loss but rather simply a transfer from the briber to the bribee.

[7] E. H. Chamberlin, "Product Heterogeneity and Public Policy," *American Economic Review,* Vol. 40, May 1950.

[8] See Richard Schmalensee, "Industrial Economics: An Overview," *Economic Journal,* September 1988, for a survey of this issue.

[9] H. Leibenstein, "Allocative Efficiency vs. X-Inefficiency," *American Economic Review,* June 1966.

[10] The pioneering work on rent-seeking behavior is Gordon Tullock, "The Welfare Costs of Tariffs, Monopolies and Theft," *Western Economic Journal* 5 (1967): 224–32. A more relevant piece for our analysis is Richard A. Posner, "The Social Costs of Monopoly and Regulation," *Journal of Political Economy* 83 (August 1975): 807–27.

[11] William P. Rogerson, "The Social Costs of Monopoly and Regulation: A Game-Theoretic Analysis," *Bell Journal of Economics* 13 (Autumn 1982): 391–401.

However, one could take the rent-seeking argument one step further and argue that agents will compete to become legislators or regulators in order to receive the rents from bribes. If real resources are used at that stage, then they represent a welfare loss.

Rent-seeking behavior can also arise in the form of excessive nonprice competition. Suppose firms are able to collude so that price exceeds cost. The lure of this high price-cost margin could generate intensive advertising competition as firms compete for market share. Depending on the particular setting, this advertising may have little social value and simply be the by-product of competition for rents. Historically, socially wasteful advertising has been thought to be a feature of the cigarette industry. As we will see in later chapters, nonprice rivalry among firms in a cartel or in a regulated industry can lead to excessive spending on product quality, product variety, and capacity as well as advertising.

Finally, unions have been found to be quite effective in extracting some of a firm's profits in the form of higher wages. This higher wage results in the private marginal cost of labor exceeding its social marginal cost, so that a firm tends to use too little labor in the production process. This inefficient input mix represents yet another source of welfare loss associated with monopoly. One study found that unions extract in excess of 70 percent of monopoly rents.[12]

Estimates of the Welfare Loss from Monopoly

Having identified various sources of welfare losses due to price exceeding marginal cost, it is natural to wonder about the quantitative size of these losses in the U.S. economy. One method for estimating the traditional deadweight welfare loss (which we will denote DWL) is as follows. From Figure 4.2, we know that DWL equals BCD when the monopoly price is charged. BCD can be approximated by $\frac{1}{2}(P_m - P_c)(Q_c - Q_m)$ where this approximation is exact if the demand function happens to be linear. More generally, if P^* is the price that firms charge and Q^* is the resulting level of demand, then DWL is approximated by $\frac{1}{2}(P^* - P_c)(Q_c - Q^*)$. Because P^* and Q^* are the actual price and quantity, one can collect data on P^* and Q^* for various firms or industries. However, we typically do not know the competitive price without estimating marginal cost. It is difficult to get a reliable estimate of marginal cost for just a single industry. To do so for a significant portion of the U.S. economy would be a gargantuan task. We then need to find some alternative way of estimating DWL that does not require having data on P_c and Q_c.

In his pioneering study, Arnold Harberger used the following approach.[13] To begin, one can perform a few algebraic manipulations and show that

$$\frac{1}{2}(P^* - P_c)(Q_c - Q^*) = \frac{1}{2}\eta d^2 P^* Q^* \tag{4.1}$$

[12] Michael A. Salinger, "Tobin's *q*, Unionization, and the Concentration-Profits Relationship," *Rand Journal of Economics* 15 (Summer 1984): 159–70.

[13] Arnold C. Harberger, "Monopoly and Resource Allocation," *American Economic Review* 44 (1954): 77–87.

where η is the absolute value of the market demand elasticity and d is the price-cost margin. More formally, $d = (P^* - P_c)/P^*$ and $\eta = |(\Delta Q/Q)/(\Delta P/P)|$ where $\Delta Q = Q_c - Q^*$ and $\Delta P = P^* - P_c$. Although data on industry revenue, P^*Q^*, are available, one needs to come up with estimates of d and η. In order to derive a ballpark figure of DWL, Harberger used the difference between an industry's rate of return and the average for the sample to estimate the price-cost margin d, and simply assumed that $\eta = 1$. With this back-of-the-envelope technique, Harberger found that DWL was on the order of one-tenth of 1 percent of GNP. Though the assumption of unit elasticity is arbitrary, what is important is that the conclusion one draws from this estimate is robust to the value of η. Even increasing it fivefold will mean that DWL is only one-half of 1 percent of GNP. Harberger concluded that the welfare losses from monopoly are very small indeed.

We thought it worthwhile to review Harberger's work in order to show how one might go about estimating welfare losses from monopoly. However, there are several reasons to question the relevance and accuracy of his low estimate of DWL. First, it is an estimate based on data from the 1920s. Whether such an estimate is relevant to today's economy is questionable. Second, we know that there are sources of welfare loss from monopoly other than DWL. Harberger estimated that the size of above-normal profits was around 3–4 percent of GNP. This leaves open the question of how much resources were used in competing for these rents. Depending on the extent of such competition, we know that the true welfare loss could be as high as 3–4 percent of GNP. The third and perhaps most important reason for questioning the validity of Harberger's estimate is that later researchers have performed more careful analyses and found higher values of DWL.

One such study was performed by Keith Cowling and Dennis Mueller.[14] They took a quite different approach to estimating DWL. Their approach avoided having to make an arbitrary assumption on the demand elasticity by assuming that firms maximize profit. The first step in their analysis is to note that a firm's profit-maximizing price P^* satisfies the following relationship:

$$\frac{P^*}{P^* - MC} = \eta \tag{4.2}$$

where MC is marginal cost. In words, a firm sets price so that the inverse of the price-cost margin equals the firm demand elasticity. Note that in a competitive industry η is infinity so that (4.2) tells us that $P^* = MC$. Recall that Harberger showed that DWL could be estimated by $\frac{1}{2}\eta d^2 P^* Q^*$ where $d = (P^* - MC)/P^*$ (and we have replaced P_c with MC). Because $1/d = P^*/(P^* - MC)$ and given (4.2), it follows that $\eta = 1/d$. Now substitute $1/d$ for η in the expression that estimates DWL [see equation (4.1)]:

[14] Keith Cowling and Dennis C. Mueller, "The Social Costs of Monopoly Power," *Economic Journal* 88 (December 1978): 727–48. For a summary of many of these studies, see Paul Ferguson, *Industrial Economics: Issues and Perspectives* (London: Macmillan, 1988).

$$DWL \cong \frac{1}{2}\eta d^2 P^* Q^* = \frac{1}{2}\left(\frac{1}{d}\right) d^2 P^* Q^* = \frac{1}{2} d P^* Q^*. \tag{4.3}$$

Substituting $(P^* - MC)/P^*$ for d in equation (4.3), it follows that

$$DWL \cong \frac{1}{2}\left(\frac{P^* - MC}{P^*}\right) P^* Q^* = \frac{1}{2}(P^* - MC)Q^* = \frac{1}{2}\Pi^* \tag{4.4}$$

where Π^* is firm profits. Because $\Pi^* = (P^* - AC)Q^*$, where AC is average cost, the last equality in (4.4) uses the assumption that marginal cost is constant so that $MC = AC$. Cowling and Mueller showed that the deadweight welfare loss created by a firm is approximately equal to half of its profits.

With this methodology, Cowling and Mueller collected data on Π^* for 734 U.S. firms for 1963–1966. Remember that Π^* represents *economic* profits, not *accounting* profits. Hence they used 12 percent as the normal return on capital in the economy and subtracted normal profits from accounting profits to estimate Π^*. Their estimate of DWL was around 4 percent of GNP, considerably higher than that found by Harberger. If one includes advertising expenditures as wasted resources associated with rent-seeking behavior, their measure jumps to 13 percent of GNP. Of course, inclusion of all advertising expenditures assumes that all advertising lacks any social value. This assumption is clearly false, because some advertising reduces search costs for consumers. Thus one would expect Cowling and Mueller's best measure of the welfare loss from monopoly to lie somewhere between 4 and 13 percent of GNP. Nevertheless, it is interesting that under their most comprehensive measure, General Motors by itself created a welfare loss of one-fourth of 1 percent of GNP!

It is clearly important to understand the quantitative size of the welfare loss from price exceeding marginal cost, whether it is due to monopoly, collusion, or regulation. Unfortunately, estimating welfare losses is an inherently precarious task because of data limitations. One must then interpret these estimates with considerable caution. A final point is that even if we knew for certain that monopoly welfare losses were, say, only 1 percent of GNP, this would not be grounds for abolishing antitrust. The reason is that the 1 percent figure would apply to an economy with antitrust in place. Perhaps if antitrust did not exist, the monopoly losses would be much larger.

Technical Progress

Efficiency in producing the desired bundle of known goods and services with a given technology is obviously important. Some argue, however, that economists place too much emphasis on this type of efficiency. They believe it is at least as important for industry to be efficient in generating new knowledge that saves resources in producing known products, as

well as in creating new or higher-quality products. In short, industry should be technically progressive.

Importance of Technological Change

In a path-breaking 1957 study,[15] Nobel laureate Robert M. Solow of MIT estimated that about 80 percent of the increase in gross output per worker-hour from 1909 to 1949 in the United States could be attributed to technological change. Subsequent studies[16] have led to somewhat lower estimates, but Solow's general conclusion as to the relative importance of technological advance is unchanged. It should be useful to illustrate his analysis graphically in order to clarify the meaning of technological change.

In Figure 4.5, two production functions are shown. The functions apply to the economy as a whole and show that output per worker-hour, Q, rises (at a decreasing rate) with the amount of capital per worker-hour, K. The lower production function represents the best technology known at time $t = 1$. New knowledge at time $t = 2$ leads to a shift upward in the function, enabling society to obtain higher Q for any given K. Thus the shift represents technological change between $t = 1$ and $t = 2$.

We can now indicate Solow's method of analysis. Suppose that at $t = 1$ the amount of capital per worker-hour is K_1 and at $t = 2$ it is K_2. Furthermore, suppose that Q_1 and Q_2 are the observed outputs per worker-hour on these two dates. The total increase in Q can be conceived as consisting of two parts: the movement from A to B (the effect of technological change) and the movement along the production function from B to C (the effect of increased capital per worker-hour). As stated earlier, Solow found that the amount of the total increase in Q due to technological change (the movement from A to B) was greater than that due to increased capital per worker-hour (the movement from B to C).

The importance of new products is also clear. One has only to think of some examples: jet aircraft, VCRs, antibiotics, personal computers, nuclear power, and so forth. This dimension of technological change was not incorporated fully in Solow's estimates.

Granted that technological change is important, we must now consider what determines it. At the industry level, it is reasonable to expect a number of factors to be influential in determining the rate of technical advance. Undoubtedly, the amount of resources devoted to research and development (R & D) is important. But the amount of private resources allocated will depend upon profitability considerations, which, in turn, will depend on such things as the expected demand for the product and the technical feasibility of the project. And, what is particularly relevant in this book, the structure of the market should affect these profitability calculations.

[15] R. M. Solow, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957.

[16] E. F. Denison, *Trends in American Economic Growth*, 1929–1982 (Washington, D.C.: Brookings Institution, 1985).

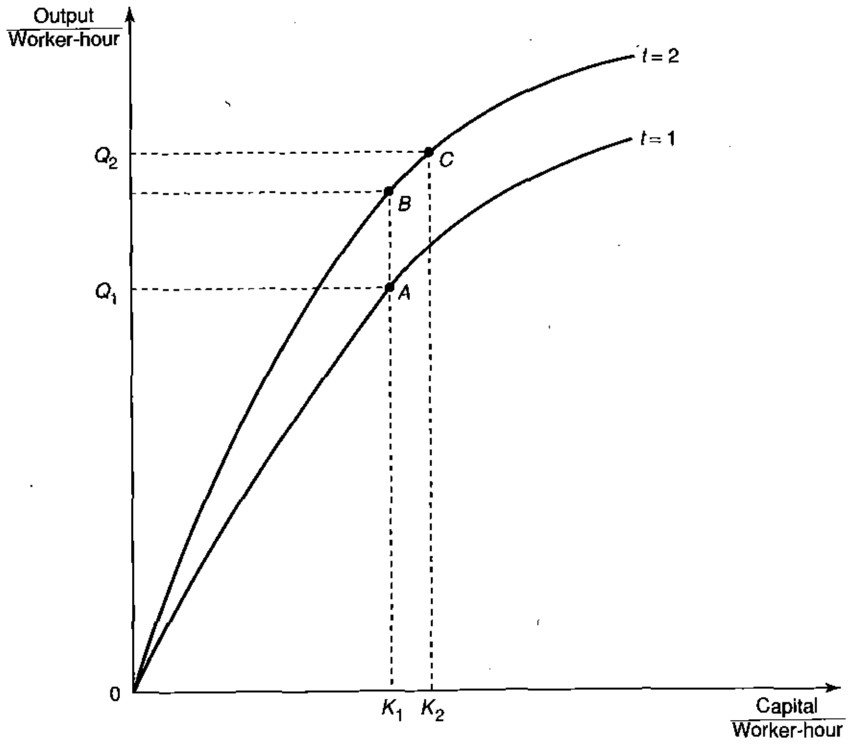

Figure 4.5
Technical Change Shifts the Production Function

Some quite persuasive economists have argued that some monopoly power is necessary to provide incentives for firms to undertake research and development programs. The rationale for existing patent policy rests to some extent upon this argument. Others, however, have taken the opposite position, namely, that it is competitive pressures that produce the higher rates of progressiveness.

The famous economist Joseph Schumpeter is usually credited with the view that some monopoly must be tolerated to obtain progressiveness. According to Schumpeter:

> But in capitalist reality as distinguished from its textbook picture, it is not [perfect] competition which counts, but the competition from the new commodity, the new technology, the new source of supply, the new type of organization . . . —competition which strikes not at the margins of the profits and the outputs of the existing firms but at their very foundations and their very lives.[17]

[17] Joseph A. Schumpeter, *Capitalism, Socialism, and Democracy* (New York: Harper & Row, 1975), p. 84.

Before turning to a rivalry model that provides some insight into these issues, it may be helpful to explain several terms that will be used in our discussions. At the beginning there is *basic research,* which seeks knowledge for its own sake. Most industrial firms engage in *applied research,* which is directed toward a particular product or process. If successful, *invention* takes place, which is the discovery of new knowledge. After invention, *development* must take place, leading to the commercial application of the invention, or *innovation.* The last phase of technical change is the *diffusion* of the product or process throughout the industry, or economy.

An R & D Rivalry Model

F. M. Scherer and D. Ross have presented an instructive model of R & D rivalry in their book *Industrial Market Structure and Economic Performance.* Their model is useful in illuminating the conflicting incentives that market structure provides for innovation: (1) more rivals tend to stimulate more rapid innovation in order to be first with a new product and benefit from the disproportionate rewards of being first, and (2) more rivals split the potential benefits into more parts, making each firm's share less. Here we shall draw heavily on their expositional approach, which, in turn, is an attempt to simplify more mathematically complex models published elsewhere.

The model collapses innovative activity into a determination of the speed of new product development. That is, the model seeks to show what factors lead to the firm's choice of the number of years from beginning R & D to the market introduction of the product. We should note that it is incorrect to equate a shorter time necessarily with "socially preferred." While we often seem to identify higher rates of innovation as necessarily "good," it is of course possible for innovation to take place too rapidly.[18]

The situation is one of oligopoly with each firm competing through improved products. To improve one's product requires carrying out R & D for a certain time period prior to marketing. The time period can be compressed by expending more resources. Hence there is a cost-time trade-off that is shown in Figure 4.6 as the curve CC'.

It is easy to explain the curve CC' by example. Let one plan be to spend $400,000 per year for 10 years. The present discounted value of this stream at 10 percent is $2.5 million. Hence this value is one point on CC'. Another plan is to spend $1 million per year for 5 years—with a present value of $3.8 million. This is a second point on CC'. Clearly the implication is that it costs more to shorten the time to innovation. There are several reasons for this: Costly errors can be made when development steps are taken concurrently instead of waiting for the information early experiments supply. Second, parallel experimental approaches may be necessary to hedge against uncertainty. Third, there are diminishing returns

[18] See, for example, Yoram Barzel, "Optimal Timing of Innovation," *Review of Economics and Statistics,* August 1968.

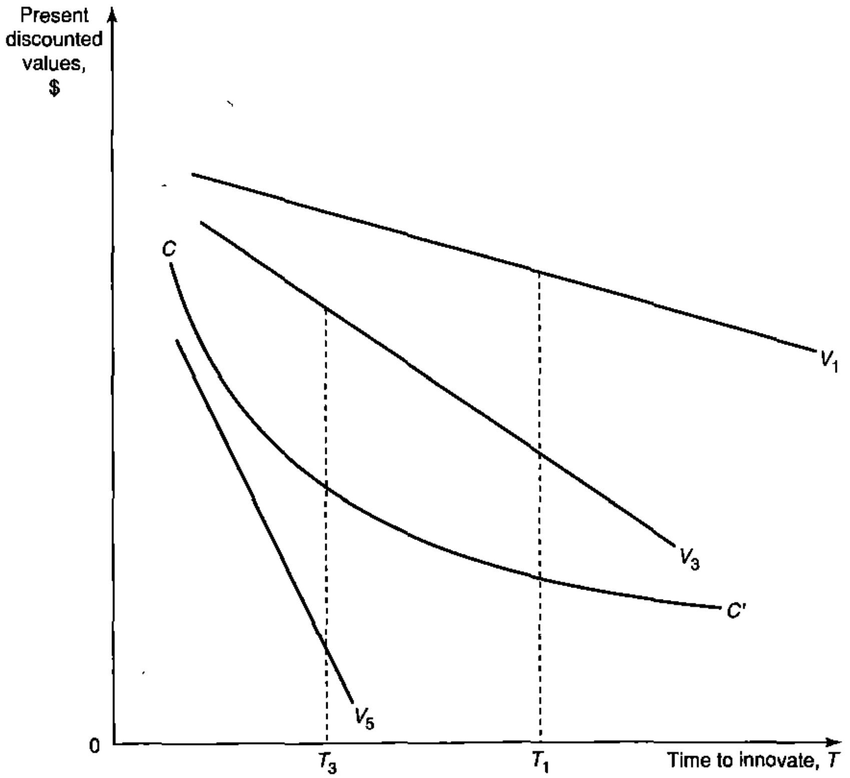

Figure 4.6
R & D Rivalry

in the application of additional scientific and engineering manpower to a given technical project.[19]

It is assumed that firms choose the time to innovation T in order to maximize the present discounted value of their profits. Hence the next step is to introduce the function V, which represents how the present value of net revenues varies with T. The net revenues are equal to revenues from the sale of the product minus the production and marketing costs incurred. As shown in Figure 4.6 the V functions (each V function corresponds to a different number of rivals) slope down to the right. It is easy to explain the slope of V_1, which refers to a monopoly situation with no rivals.

[19] F. M. Scherer and D. Ross, *Industrial Market Structure and Economic Performance*, 3rd ed. (Boston: Houghton Mifflin, 1990), p. 632.

Assume for simplicity that the net revenues from the product will be constant over time. Now, if somehow T happened to be zero, the vertical intercept of V_1 would equal the present value of this constant stream of net revenues from $T = 0$ forever. If the flow is $1 million per year, then the present value at 10 percent would be $10 million. Now as T increases, the early years of potential net revenues are lost, thereby reducing the present value and causing the V_1 function to slope down to the right. For example, if net revenues do not begin until year 3, the present value falls to $8.3 million.

In this monopoly case, the profit-maximizing T is easily found graphically. It is simply that value of T that is associated with the largest vertical difference between the present value of net revenues and the present value of R & D costs. This is also found by locating the value of T where the slope of V_1 equals the slope of CC'. The optimal T is shown as T_1 in the figure.

Now consider a second situation in which there are, say, three rivals. This is represented with the function V_3. Two points should be noted about V_3 relative to V_1. It is lower, reflecting lower net revenues for each T, and it is steeper. Thus, V_3 is lower than V_1 simply because the total market potential net revenues must now be split three ways. That is, it is reasonable for a firm with two rivals to expect to share the market with the other two, to some degree. Notice that this shift downward reduces overall expected profits, but it does not eliminate them because V_3 still lies above CC'. This reduced expected appropriability of net revenues by the firm can lead to a situation in which the innovation is simply unprofitable—with a zero rate of innovation. Such a case is shown by the function V_5, which corresponds to five rivals. Presumably five rivals is "too many" and would result in too much imitation for R & D to be undertaken at all.

Return to the V_3 case and consider the second point made in the preceding paragraph. We see that V_3 is steeper than V_1. First note what this steeper slope implies about the optimal T. As the slope gets steeper, the optimal T falls until the V_3 function's slope equals that of CC', at T_3. This steepness, in other words, leads to a faster speed of development as compared to the monopoly case. This effect of increasing the number of rivals is therefore a stimulating effect on the rate of innovation—as long as the number of rivals does not increase too much and cause a situation where innovation is completely unprofitable.

What causes the slope of V_3 to be steeper than V_1 can be explained as follows. The idea is that the proportionate payoff to being first, and enjoying the whole market until imitation, grows with the number of rivals. In monopoly, there is little loss as one innovates later and later—the monopoly still has the whole market in later years. This means the slope of V_1 is relatively flat. Now in a three-firm market, the first firm enjoys the whole market until imitation occurs. Let us say that when imitation occurs, the leader's share falls to one-third—equal to each of the two imitators. The relative size of the leader's payoff to one of the two imitators' payoffs is what determines the slope of V_3. Clearly the relative payoff for a low T (and being first) is greater than the case of monopoly. Furthermore, in some cases the pioneer firm is even relatively better off because of brand loyalty developed during the early years. This makes it possible to keep a proportionately greater share of the market than its imitators. For example, brand loyalty may make it possible for the pioneer to keep half the market, with each imitator getting one-fourth.

Hence the model that we have described points clearly to the influence of market structure on innovation. Though the complexity of the innovative process makes it difficult to obtain nice, neat results, one can infer that neither pole of perfect competition nor pure monopoly seems to be ideal. As Scherer and Ross put it in summarizing an extensive review of empirical work:

> What is needed for rapid technical progress is a subtle blend of competition and monopoly, with more emphasis in general on the former than the latter, and with the role of monopolistic elements diminishing when rich technological opportunities exist.[20]

A more fundamental issue is that it may be naive to conceive of the public policy issue as one of choosing the optimal market structure to optimize the trade-off between static allocative efficiency and progressiveness. The reason is that structure itself should perhaps be viewed as evolving endogenously as technological change occurs through time. Thus, firms that are successful in the innovation game will grow while others decline or drop out. And, over time, the industry's concentration will change as a result.

In Chapter 24 we consider a special policy toward technological change—the granting of patents to provide incentives for inventive activity. Although the model of R & D rivalry implicitly assumed patents to be unimportant, Chapter 24 goes to the other extreme and assumes that patents are essential. Most empirical studies conclude that the importance of patents varies greatly across industries, being especially important in pharmaceuticals and chemicals.

Summary

This chapter has examined two dimensions of economic performance: efficiency and technical progress. The major difference is that the efficiency section assumed a known technology while the technical progress discussion focused on the allocation of resources to develop new knowledge (for producing new products, and for producing existing products more cheaply).

An important lesson that this chapter tries to teach is the usefulness of total economic surplus in assessing public policies. That is, if total economic surplus rises as a result of a policy change, then under certain plausible assumptions, one can argue that the change is in the public interest. An example of such a change that was described was the decontrol of oil prices in the United States.

A hypothetical monopoly-versus-competition example was used to explain the concept of the deadweight loss caused by monopoly pricing. A short section discussed several empirical studies that have sought to estimate the social cost of monopoly in the United States.

In the technical progress section, a simple model of R & D rivalry was presented. The model illustrated how increasing the number of rivals can have two opposing effects on the speed of innovation. The key point of the model is that no simple relationship between the number of rivals and the rates of innovation exists—a larger number of rivals does not always produce better results for society.

Questions and Problems

1. Explain the difference between the Pareto criterion and the compensation principle as rules for deciding whether a particular policy change is in the public interest.

2. Assume, in the monopoly-versus-competition example in the text where demand is $Q = 100 - P$ and marginal cost MC = average cost AC = \$20, that MC under competition remains at \$20. However, assume that the reason the monopoly can continue to be a monopoly is that it pays \$10 per unit of output to reimburse lobbyists for their efforts in persuading legislators to keep the monopoly insulated from competition. For example, the lobbyists may be generating (false) studies that demonstrate that competition results in higher costs.

   a. Calculate the prices and quantities under monopoly and competition.

   b. Calculate total economic surplus under monopoly and competition. The difference is the social cost of monopoly.

   c. The social cost of monopoly can be disaggregated into two distinct types of cost: the resources cost of rent seeking and the usual deadweight loss of output restriction. What are their respective magnitudes?

3. Discuss the concept of "equity cost" used in the oil industry study by Arrow and Kalt. Do you think it is generally true that "consumers" have lower incomes than "producers"? Does it matter to your answer that labor unions and senior citizens have large ownership interests in corporations through pension funds?

4. A (mini-) refrigerator monopolist, because of strong scale economies, would charge a price of \$120 and sell forty-five refrigerators in Iceland. Its average cost would be \$60. On the other hand, the Iceland Planning Commission has determined that five refrigerator suppliers would be sufficiently competitive to bring price into equality with average cost. The five-firm equilibrium would yield a price of \$100 and a total output of fifty refrigerators.

   a. Consumer surplus under the five-firm industry organization would be larger than under monopoly. If the demand curve is linear, by how much is consumer surplus larger?

   b. Producer surplus under monopoly is larger—by how much?

   c. If the Planning Commission thinks that total economic surplus is the correct criterion, which organization of the refrigerator industry will they choose?

[20] Ibid., p. 660.

5. What is the best market structure for promoting technical progress?
6. A study in 1975 estimated the effect of monopoly on equity as opposed to efficiency (W. Comanor and R. Smiley, "Monopoly and the Distribution of Wealth," *Quarterly Journal of Economics*, May 1975). For 1962, the wealthiest 0.27 percent of the population accounted for 18.5 percent of wealth. If all industries were competitive, this study estimated that the wealthiest 0.27 percent would have only 13 percent of wealth in 1962. Can you explain this finding? Hint: The wealthiest 0.27 percent held 30 percent of business ownership claims.

5 Oligopoly, Collusion, and Antitrust

Section 1 of the Sherman Act prohibits contracts, combinations, and conspiracies that restrain trade. Although this is rather general language, it usually refers to conspiracies to fix prices or share markets. In this chapter, we will trace major judicial decisions from the passage of the Sherman Act in 1890 to the present to show the evolution of the current legal rules toward price fixing.

Before beginning this task, however, we shall discuss the theories of collusive and oligopoly pricing. Oligopoly, of course, refers to a market structure with a small number of sellers—small enough to require each seller to take into account its rivals' current actions and likely future responses to its actions. Price-fixing conspiracies, or cartels, are not limited to a small number of sellers, although it is generally believed that the effectiveness of a cartel is greater when the number of participants is small.

Our coverage will proceed in the following manner. In order to explore the theory of oligopoly and collusion, we will need to be properly tooled. Toward this end, an introductory discussion of game theory is provided. With that under our belts, the plan is to review the Cournot model and a model of collusive behavior. The last section of this chapter discusses antitrust law and landmark price-fixing cases.

A very important assumption that underlies the analysis in this chapter is that potential entry is not a problem. We shall always assume that the number of active firms is fixed. Our focus is then upon the internal industry problems of firms reaching an equilibrium when the only competition comes from existing firms. Allowing for competition from new or potential entrants is delayed until the next chapter.

Game Theory

Example 1: Advertising Competition

Consider a duopoly in which firms do not compete in price because of collusion or regulation. Let the price be $15 and the quantity demanded be 100 units. If unit cost is $5, then profit per unit equals $10. That is, a firm receives revenue of $15 for each unit, and it costs the firm $5 to produce that unit. Though it is assumed that firms have somehow been able to avoid competing in price, it is also assumed that firms do compete via advertising. To simplify matters, a firm can advertise at a low rate (which costs $100) or at a high rate (which costs $200). Also for simplicity, assume that advertising does not affect market demand but rather just a firm's market share. Specifically, a firm's market share depends on how much it advertises relative to its competitor. If both firms advertise an equal amount (whether low or high), then firms equally share market demand—that is, each has demand of 50 units. However, if one firm advertises low and the other advertises high, then the high advertising firm dominates the market with a market share of 75 percent.