In other words, the maximum price that the monopolist can set in the first period in order to sell just one unit is $(1+R)\left(\$30+\dfrac{\$20}{1+R}\right)=(1+R)\$30+\$20$.

In light of the above, we can now greatly simplify the monopolist's problem. Either sell two units in the first period each at a price of $(1+R)\$30$; or one unit in the first period at a price of $(1+R)\$30+\$20$ and a second unit in the following period at a price of \$30; or no units in the first period but both units in the second period at a price of \$30. The corresponding profits are shown in Table 2.1.

A little algebra will confirm that the first row option in Table 2.1 gives the greatest profit in present value terms. That is, the monopolist will maximize profit by setting a first-period price of $(1+R)\$30$ and selling both units immediately, one to each consumer. The monopolist will then earn a surplus of $2(1+R)\$30$. The high-value consumer will earn a surplus of $(1+R)\$20$, while the low-value consumer will earn zero surplus. The total surplus will thus be $2(1+R)\$30+(1+R)\$20=(1+R)\$80$, the maximum amount possible. In other words, the durable feature of the good has constrained the monopolist to set efficient prices—ones that maximize the surplus. This is Coase's conjecture. It is the supposition that durability constrains monopoly pricing so as to eliminate any deadweight loss.

However, just as the efficacy of entry as a discipline on monopoly behavior can be weakened, so too can the impact of durability. Consider how the above example would have been altered had the low-value consumer only valued the services of the durable good at \$20 per period or $(1+R)\$20$ in total. In this case, it is now no longer the case that any and all second-period sales must take place at the maximum price of the low-value consumer, which is now \$20. If the monopolist sells no units in the first-period and enters the second-period with two units to sell, it will be preferable to sell just one unit to the high-value consumer at \$50 rather than both units at \$20 a piece. It follows that the high-value consumer can no longer bargain so aggressively in the first period. This consumer will now prefer any price of $(1+R)\$(50-\varepsilon)$ in the first period that provides a surplus of $(1+R)\$\varepsilon$ to a price of \$50 in the second-period that provides a surplus of zero. Thus, by making ε arbitrarily small, the monopolist can induce the high-value consumer to buy in the first period, and then sell at a price of \$20 to the low-value consumer in the second period. Neither consumer will enjoy any surplus, while the monopolist will earn a profit of $(1+R)\$50+R\$20$. As can be easily verified, the maximum surplus available in this case is $(1+R)\$70$, which is \$20 greater than the surplus actually realized. Here, durability has not eliminated the deadweight loss of monopoly.

The difference in the two cases reflects the distribution of consumer valuations. In both cases, this distribution is discrete. However, the jump from one valuation to the next is more substantial in the second case. Heuristically, valuations are less continuous in that setting. This points to a generalization regarding the Coase conjecture. The more that

Table 2.1 Price options and corresponding profit for the durable goods monopolist

First-Period Price	*Second-Period Price*	*Present Value of Profits*
$(1+R)\$30$	NA	$2(1+R)\$30$
$(1+R)\$30+\$20$	\$30	$(1+R)\$30+\$20+R\$30$
$>(1+R)\$50$	\$30	$2R\$30$

consumer valuations are continuous, the more that durability constrains monopoly pricing of durable goods to be efficient. Indeed, as the distribution becomes perfectly continuous, the monopolist may be forced to set competitive prices. In general, durability will, like potential entry, impose some limits on the exploitation of monopoly power and the extent of the associated welfare loss.[13]

2.3.4 The Nonsurplus Approach to Economic Efficiency[14]

Let us return to the second example in the discussion of the Coase conjecture above in which the Coase conjecture fails. Recall that the efficient first-period price is $(1 + R)\$20$ resulting in a total surplus of $(1 + R)\$70$, but that the monopolist instead sets an initial price of $(1 + R)\$50$ and a subsequent second-period price of \$20, resulting in a deadweight loss of \$20. We may ask now how matters might be different had there been fifty high-value consumers and fifty low-value ones again assuming that the monopolist has just two units.

The first thing to note is that the total surplus now available on the market is $2(1 + R)\$50$. At a first-period price of $(1 + R)\$50$, two units will be sold to two of the 50 high-value consumers and all of this surplus will go to the monopolist. In fact, no other price is possible. None would buy at a higher price. Moreover, at any lower price, the high-value consumers would actively compete with higher bids until again the price rose to $(1 + R)\$50$. Here is another case in which profit maximization by the monopolist is consistent with efficient resource allocation.

Why does this happen here? Why is it that now, despite the sizeable differences in consumer valuations between the two groups, monopoly pricing involves no deadweight loss? The answer lies in considering more deeply just what that deadweight loss of monopoly signifies. The deadweight loss of monopoly is really just the additional surplus that the monopolist could generate if output were increased to the efficient level. The problem for the monopolist in the conventional setting, e.g., Figure 2.2(b), is that the monopolist cannot appropriate the full value of this marginal surplus. As the additional output is sold and price moves down the demand curve, units are sold at a lower price. Even if the monopolist captures all the surplus on the marginal units, the net gain is less than this because less is now earned on infra-marginal units.

It is the inability of the monopolist to appropriate all the surplus that production generates that lies at the heart of the welfare loss of monopoly. When there are fifty high-value consumers, however, and the monopolist only has two units to sell, this problem no longer exists. The monopolist can sell the efficient amount without lowering the price to any infra-marginal consumer. Within the feasible range of sales, the monopolist generates $(1 + R)\$50$ of surplus for each unit sold and acquires and keeps all this surplus.

When there are fifty high-value consumers, the monopolist's output potential is small relative to the total market. At the margin, the decision to increase or decrease sales by one unit and sell it at $(1 + R)\$50$, leaves the surplus of all other market participants unaffected. The marginal consumer earns a surplus of zero in either case. Note that this is similar to the perfectly competitive case. There, too, each firm is sufficiently small relative to the market that it has no effect on the surplus earned by all other participants. The profit of a

[13] Stokey (1981) offers a formal derivation of the Coase Conjecture. See also Thepot (1998).

[14] This section and the previous one make extensive use of the nonsurplus approach developed in Makowski and Ostroy (1995). It has had an important influence on our understanding of market participation. It also plays a central role in the business strategies advocated by Brandenburger and Nalebuff (1996).

single competitive firm at the margin is zero and this is precisely the same as that firm's contribution to the surplus or welfare created by market trading. The same can be true for a monopolist if it is small relative to the overall market.

The understanding that the crucial factor giving rise to inefficient outcomes is that there are market participants able, at the margin, to alter the surplus of other participants is known as the non-surplus approach to economic analysis. For our purposes, it makes the important point that it is not the number of sellers but their size relative to the market that gives rise to a welfare loss. In other words, it is not the fact that the monopolist is the only seller but the fact that the monopolist is large relative to the market that leads to inefficient outcomes.[15]

2.4 EMPIRICAL APPLICATION: TESTING THE COASE DURABLE GOODS MODEL

Coase's argument that monopolists selling durable goods may still face market pressure to price competitively rests on a very rational model of consumer behavior in which consumers are assumed to be forward-looking when making today's buying decisions. Whether buyers are really that rational, though, is open to question. Judith Chevalier and Austan Goolsbee (2009) examine this issue by analyzing the college textbook market. Their intuition is straightforward: If the standard model accurately describes consumer behavior, then students buying a textbook should consider the likelihood that a new edition will come out after the student is done with it and wants to sell it back on the used book market. In particular, the prospect that a new edition will come out soon should depress the student demand for the book, as this will make it harder to sell a used version.

To investigate this issue, Chevalier and Goolsbee collected data for textbooks in biology, economics, and psychology for the years 1997 through 2001 by semester and by class. For each class, they know the assigned textbook and how many people were in the class. They also know the aggregate new book sales of each text and also the total sales of used versions of the assigned text through college bookstores. For any textbook i in year t, let N_{it} be the number of students who bought the text new and A_{it} be the total number of students in classes in which the book was assigned. Then $S1_{it} = N_{it}/A_{it}$ is the share of the potential academic class market that bought the book new. Similarly, let $S2_{it}$ be the share of students who bought a used version at the college bookstore. Finally, let $S3_{it}$ be the share of students who did not buy the assigned text at all. The variable $S1_{it}/S3_{it}$ is then a measure of the demand for a new, unused book relative to not buying the book in any form. It is tempting to relate this linearly to the price p_{it} of a new book and some other independent variables X_{it}, to give the simple linear:

$$S1_{it}/S3_{it} = \beta \mathrm{X}_{it} - \alpha_1 p_{it} + u_{it} \tag{2.7}$$

Here, β is a vector of coefficients on the X_{it} variables, which include such things as whether the book is an introductory or advanced text, whether it comes shrink-wrapped with a study guide and other ancillaries, and the book's level of difficulty, while α_1 measures the effect of the book price on student purchases of new, unused texts. However, Chevalier

15 The nonsurplus approach is developed in Makowski and Ostroy (1995). It has had an important influence on our understanding of market participation. It also plays a central role in the business strategies advocated by Brandenburger and Nalebuff (1996).

and Goolsbee recognize that because both $S1_{it}$ and $S3_{it}$ are positive fractions, their ratio must also always be positive. Yet it is possible that the data may give rise to estimates of β and α_1 that predict values below zero. Further, the simple relationship expressed in equation (2.7) allows no direct test of whether students are forward-looking in their textbook purchases.

Chevalier and Goolsbee (2009) therefore make two alterations to the simple model above. First, they make what is known as a logit transformation by using the natural log of the dependent variable, i.e., by using $\ln(S1_{it}/S3_{it}) = \ln(S1_{it}) - \ln(S3_{it})$. Because logarithms can be negative, this allows the data to "speak freely." Second, they include a second price variable $= .5(1 - D_{it})p_{it}$, where D_{it} is a 1, 0 dummy variable that takes the value 1 in any semester in which a new edition of the book is coming out one semester later, and zero otherwise. The idea behind this second modification is simple. The true price of buying a new textbook is the price paid for it initially, p_{it}, less the discounted price for which students can sell the book back at the end of the term. Chevalier and Goolsbee find that this resale price is typically on the order of 50 to 75 percent of the initial price, or $0.5p_{it}$, reflecting a rule of thumb for buybacks that almost all bookstores follow. However, if a new version of the text is about to come out ($D_{it} = 1$), then the resale price falls to zero. Thus, the true price of purchasing a new book is given by: $p_{it} - R0.5(1 - D_{it})p_{it}$. Here, R reflects the discount factor that students apply to the money they will receive in the future when they sell the book—a variable that is not directly observed.

With these modifications, equation (2.7) becomes:

$$\ln(S1_{it}) - \ln(S3_{it}) = \beta \mathrm{X}_{it} - \alpha_1[p_{it} - R0.5(1 - D_{it})p_{it}] + u_{it} \tag{2.8}$$

or, more generally,

$$\ln(S1_{it}) - \ln(S3_{it}) = \beta \mathrm{X}_{it} - \alpha_1 p_{it} + \alpha_2(1 - D_{it})p_{it} + u_{it} \tag{2.9}$$

Note that the value of $\alpha_2 = \alpha_1 0.5R$ and that it indicates whether or not students are forward-looking. If $\alpha_2 = 0$, then students are myopic and consider only the initial price. However, if α_2 is not zero, then students do consider a text's future resale value in their original purchase decision, except in years when $D_{it} = 1$ so that future resale value is zero. Note as well that we can define the ratio $\lambda = \alpha_2/\alpha_1 = 0.5R$. Hence, if we have estimates of α_1 and α_2, we can infer the typical student's discount factor R.

Chevalier and Goolsbee (2009) estimate equation (2.9) with alternative regression techniques (OLS and GMM) including in each case several additional independent variables (X_{it}). These are time dummies to capture general shifts in textbook demand from one year to the next, for the particular field that the text is in, whether it is bundled with additional software; the fraction of the time that the book was required for the class and not just assigned (which is greater than 0.90 for all of the books considered); the average SAT score of students assigned the book; the age of the edition; and whether it is a paperback.[16] For the purposes of testing the Coase view and, specifically, the underlying assumption that

[16] Students with a knowledge of econometrics may recognize that the price variable is endogenous and that therefore an instrumental variables (IV) approach is warranted. Chevalier and Goolsbee do indeed adopt an IV approach, including a dummy variable indicating if a book is published by a non-profit publisher, the share of non-profit publishers among textbooks designed for the same course, and the Herfindahl index for publishers for the course in the year in which the textbook was published.

Table 2.2 Estimated effects of current price and future resale price on textbook demand (standard errors in parentheses)

Coefficient	*OLS Estimation*	*GMM Estimation*
α_1	−0.060 (0.008)	−0.061 (0.012)
α_2	0.033 (0.003)	0.037 (0.003)
λ	0.55 (0.092)	0.61 (0.15)

students are forward-looking, we present only their basic results for the price variables. These are shown in Table 2.2.

These results confirm both the model and the hypothesis that students are definitely forward-looking. The current price has a highly significant and negative effect on new textbook demand, while the resale price has a significantly positive effect just as it should for forward-looking consumers. Indeed, the estimates of λ imply estimated student discount factors that exceed 1. If this were literally true, it would mean that far from being myopic, students care more about the future resale value than the initial purchase price in buying a text. However, the estimate of λ is not so precise that one can rule out more normal, one-semester discount factors of say $R = 0.98$.

In short, the Chevalier and Goolsbee (2009) estimates imply that consumers, or at least college students taking biology, economics, and psychology classes, are forward-looking. Indeed, the difference in price sensitivity of these students differs dramatically in years in which a new version of the text is not anticipated ($D_{it} = 0$) and years in which one is ($D_{it} = 1$). In the former case, the estimates in Table 2.2 imply that a 10 percent rise in the initial purchase price will decrease initial sales by only 16.3 percent because they plan on getting half of that price increase back at the end of the term. In contrast, for students expecting no resale because of an anticipated new edition, a 10 percent price rise on the initial purchase leads to a nearly 40 percent decline in sales. Perhaps somewhat paradoxically, then, the presence of a resale market leads to higher initial prices as students are more willing to pay those prices if they believe that can resell the book later.

As authors of a textbook ourselves, we know that it is widely believed among students that new editions are introduced to raise profits by eliminating competition from used books. Yet Chevalier and Goolsbee (2009) show that their estimates imply that there is no such incentive for economics textbooks because given how forward-looking students are and how sensitive they are to a book's expected resale value, the expected sales revenue for economics texts varies little with the frequency of revision. Faster revisions means lower initial sales in a manner that basically cancels out any gain from eliminating used book competition. This evidence thus lends fairly strong support to Coase's argument regarding the pricing constraints facing a durable goods monopolist.

Summary

We have formally presented the basic microeconomic analysis of markets characterized by either perfect competition or perfect monopoly. In both cases, the goal of any firm is assumed to be to maximize profit. The necessary condition for profit maximization is that the firm produce at a level at which marginal revenue equals marginal cost. Because firms in competitive markets take market price as a given, price equals marginal revenue for the competitive firm. As a result, the competitive market equilibrium is one in which price is set equal to marginal cost. In turn, this implies that the competitive market equilibrium is efficient in that it maximizes the sum of producer and consumer surplus.

The pure monopoly case does not yield an efficient outcome. The monopoly firm understands that it can affect the market price and this implies that marginal revenue will be less than the price for a monopoly firm. If the market demand curve is linear, this difference is reflected in the fact that the monopolist's marginal revenue curve has the same price intercept but is twice as steeply sloped as its demand curve. A monopoly firm equating marginal revenue with marginal cost, as required for profit maximization, yields an output inefficiently below that of the competitive equilibrium. Resources are misallocated because too few resources are employed in the production of the monopolized commodity. The inefficiency that results is often called the deadweight or welfare loss of monopoly.

There are, however, some natural limits to monopoly power. To begin with, monopoly profits will attract entry so that firms with substantial market shares may be forced to price relatively close to competitive levels by potential entrants who will swoop in and steal customers away (perhaps permanently) should the incumbent firm materially abuse its market power. In addition, monopolists of durable goods face competition from still-surviving units of their own earlier production. This creates a difficult pricing tension that can again discipline the firm's pricing and, depending on the distribution of consumer reservation prices, even force it to charge perfectly competitive prices. To be sure, this discipline requires that consumers are forward-looking and that they consider future used good prices when they make their initial purchase. Evidence from the market for student textbooks tends to confirm that this is the case.

Perfect competition and monopoly form the two polar cases of market structure. The behavior of firms and the nature of market outcomes in these two cases are fairly well understood. The remaining question is what happens in the far more common cases of something between these two extremes. Before we directly address that issue, however, we need to have a more rigorous way of characterizing the market setting and a deeper understanding of the technological and cost features that give rise to particular market structures. These are the topics to be addressed in the next two chapters.

Problems

1. Suppose that the annual demand for prescription antidepressants such as Prozac, Paxil, and Zoloft is, in inverse form, given by: $P = 1000 - 0.025Q$. Suppose that the competitive supply curve is given by: $P = 150 + 0.033Q$.
   a. Calculate the equilibrium price and annual quantity of antidepressants.
   b. Calculate i) producer surplus; and ii) consumer surplus in this competitive equilibrium.

2. Assume that the dairy industry is initially in a perfectly competitive equilibrium. Assume that, in the long run, the technology is such that average cost is constant at all levels of output. Suppose that producers agree to form an association and behave as a profit-maximizing monopolist. Explain clearly in a diagram the effects on (a) market price, (b) equilibrium output, (c) economic profit, (d) consumer surplus, and (e) efficiency.

3. Suppose that the total cost of producing pizzas for the typical firm in a local town is given by: $C(q) = 2q + 2q^2$. In turn, marginal cost is given by: $MC = 2 + 4q$. (If you know calculus, you should be able to derive this expression for marginal cost.)
   a. Show that the competitive supply behavior of the typical pizza firm is described by: $q = \frac{P}{4} - \frac{1}{2}$.
   b. If there are 100 firms in the industry, each acting as a perfect competitor,

show that the market supply curve is, in inverse form, given by: $P = 2 + Q/25$.

4. Let the market demand for widgets be described by $Q = 1000 - 50P$. Suppose further that widgets can be produced at a constant average and marginal cost of \$10 per unit.
   a. Calculate the market output and price under perfect competition and under monopoly.
   b. Define the point elasticity of demand ε_D at a particular price and quantity combination as the ratio of price to quantity times the slope of the demand curve, $\Delta Q/\Delta P$, all multiplied by -1. That is, $\eta_D = -\frac{P}{Q}\frac{\Delta Q}{\Delta P}$. What is the elasticity of demand in the competitive equilibrium? What is the elasticity of demand in the monopoly equilibrium?
   c. Denote marginal cost as MC. Show that in the monopoly equilibrium, the following condition is satisfied: $\frac{P - MC}{P} = -\frac{1}{\eta_D}$.

5. Suppose that the inverse demand for clothes hangers is given by: $P = 3 - Q/16{,}000$. Suppose further that the marginal cost of producing hangers is constant at \$1.
   a. What is the equilibrium price and quantity of hangers if the market is competitive?
   b. What is the equilibrium price and quantity of hangers if the market is monopolized?
   c. What is the deadweight or welfare loss of monopoly in this market?

6. A single firm monopolizes the entire market for single-lever, ball-type faucets which it can produce at a constant average and marginal cost of $AC = MC = 10$. Originally, the firm faces a market demand curve given by $Q = 60 - P$.
   a. Calculate the profit-maximizing price and quantity for the firm. What is the firm's profit?
   b. Suppose that the market demand curve shifts outward and becomes steeper. Market demand is now described as $Q = 45 - 0.5P$. What is the firm's profit-maximizing price and quantity combination now? What is the firm's profit?
   c. Instead of the demand function assumed in part b, assume that market demand shifts outward and becomes flatter. It is described by $Q = 100 - 2P$. Now what is the firm's profit-maximizing price and quantity combination? What is the firm's profit?
   d. Graph the three different situations in parts a, b, and c. Based on what you observe, explain why there is no supply curve for a firm with monopoly power.

References

Baumol, W. J., J. C. Panzar, and R. D. Willig. 1982. *Contestable Markets and the Theory of Market Structure*. New York: Harcourt, Brace, Jovanovich.

Brandenburger, A., and B. Nalebuff. 1996. *Co-opetition*. Cambridge: Harvard University Press.

Chevalier, J., and A. Goolsbee. 2009. "Are Durable Goods Consumers Forward-Looking? Evidence from College Textbooks," *Quarterly Journal of Economics* 124 (November): 1853–84.

Coase, R. L. 1972. "Durability and Monopoly," *Journal of Law & Economics* 15 (April): 143–50.

Makowski, L., and J. Ostroy. 1995. "Appropriation and Efficiency: A Revision of the First Theorem of Welfare Economics," *American Economic Review* 85: 808–27.

Marshall, Alfred. 1890. *Principals of Economics*, London: Macmillan Publishers.

Stokey, Nancy L. 1981. "Rational Expectations and Durable Goods Pricing," *Bell Journal of Economics* 12 (Spring): 112–28.

Thepot, Jacques. 1998. "A Direct Proof of the Coase Conjecture," *Journal of Mathematical Economics* 29 (January): 57–66.

Appendix

The Calculus of Competition

The competitive firm's problem may be solved by first writing the firm's profit π as a function of its output, π(q) in turn defined as revenue R(q) less cost C(q). Price-taking behavior implies R(q) = Pq. Hence:

$$\pi(q) = \mathrm{R}(q) - C(q) = Pq - \mathrm{C}(q) \tag{2.A1}$$

Standard maximization then yields:

$$\frac{d\pi}{dq} = P - C'(q) = 0 \tag{2.A2}$$

P = marginal cost $C'(q)$ under perfect competition.

For the monopoly firm, its output is the same as industry output Q, that gives $P(Q)$, via the inverse demand curve. Hence, the monopolist's profit maximization problem is to choose output Q so as to maximize:

$$\pi(Q) = \mathrm{R}(Q) - \mathrm{C}(Q) = P(Q)Q - \mathrm{C}(Q) \tag{2.A3}$$

Again, standard maximization techniques yield:

$$\frac{d\pi}{dQ} = \mathrm{P}(Q) + Q P'(Q) - \mathrm{C}'(Q) = 0 \tag{2.A4}$$

$P(Q) + Q\mathrm{P}'(Q)$, is the firm's marginal revenue. The monopolist will maximize profit by producing where marginal cost equals marginal revenue. For a linear demand curve of the form of $P(Q) = \mathrm{A} - \mathrm{B}Q$ we have $\mathrm{P}'(Q) = -\mathrm{B}$. In this case, the firm's marginal revenue is $\mathrm{A} - \mathrm{B}Q - \mathrm{B}Q$, or $\mathrm{A} - 2B\mathrm{Q}$. The monopolist's marginal revenue curve has the same intercept as its demand curve twice the slope.

Note that the profit-maximizing condition above can also be written as

$$P(Q) - C'(Q) = -QP'(Q) \tag{2.A5}$$

Dividing both sides by $P(Q)$, we then have

$$\frac{P(Q) - C'\,(Q)}{P(Q)} = -\frac{QP'(Q)}{P(Q)} = \frac{1}{\eta} \tag{2.A6}$$

where η is what economists call the elasticity of demand—a measure of how responsive the quantity demanded is to price movements. It is formally defined as:

$$-\eta = \frac{P(Q)}{Q}\frac{1}{P'(Q)} \tag{2.A7}$$

3

Market Structure and Market Power

Early industrial economics, working in the Structure-Conduct-Performance framework, examined firms' decisions given the industrial structure. More modern analysis, however, recognizes that firms' strategic behavior will in fact be a major determinant of market structure. Yet despite this change in perspective, it is clear that contemporary industrial economics must still address the issue of market structure, or the way the industry's producers are organized. In turn, this requires that market structure—and its close cousin market power—be well understood in a way that allows distinction between different structures or different degrees of market power.

We know, for example, that markets work well when the market consists of numerous firms, each with a minimal market share. Yet such markets are relatively rare in the real world. Some markets have just two or three firms. Some have ten or twelve of unequal size. In what ways is this difference important? If there are twenty firms, does it matter if one firm has 60 percent of the market and the other nineteen have just a bit more than 2 percent each? Alternatively, can we measure market structure in such a way that enables us to make some inference of market power? Can we create an index that allows us to say how close or how far a market structure is from the competitive ideal? Because such a roadmap could be of great use to policy makers, it is worthwhile to explore the question at some length.

3.1 DESCRIBING MARKET STRUCTURE

One way to think about an industry's structure is to undertake the following simple procedure. First, take all the firms in the industry and rank them by some measure of size from largest to smallest—one, denoting largest; two, the next largest; etc. Suppose for example that we use production as a measure of size. We could then calculate the fraction of the industry's total output that is accounted for by the largest firm, then the two largest firms combined, then the three largest firms combined, and so on. This gives us the cumulative fraction of the industry's total output as we include progressively smaller firms. Plotting this relationship yields what we call a concentration curve. It is called a concentration curve because it describes the extent to which output is concentrated in the hands of just a few firms.

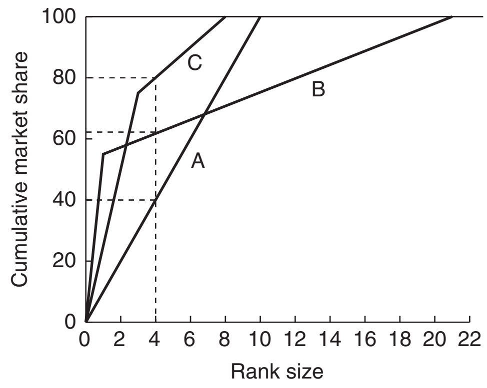

Figure 3.1 Some possible concentration curves

Figure 3.1 displays concentration curves for each of three representative industries: A, B, and C. The firms' ranked sizes are measured along the horizontal axis, again with the first firm being the largest. The cumulative market share is measured on the vertical axis. For example, Industry A has ten firms, each with a 10 percent market share. Industry B has twenty-one firms, the largest of which has a 55 percent market share. The remaining twenty firms each have a 2.25 percent share. Finally, in Industry C, there are three firms each with a market share of 25 percent and five firms each with a market share of 5 percent. For Industry B, the vertical coordinates corresponding to the horizontal values 1 and 2 on the concentration curve are 55 and 57.25, respectively. This reflects the fact that the largest firm has 55 percent of the market and the largest two firms have 57.25 percent between them.

Concentration curves are a useful illustrative device.[1] They permit one to get a quick sense of how industry production is allocated across firms from a swift visual inspection. However, often we need to summarize industrial structure with just a single parameter or index. One such measure that is frequently used is the four-firm concentration ratio, CR_4, or the percent of industry sales accounted for by the top four firms. For the three hypothetical industries described above, we can identify the CR_4 concentration very easily. All we need to do is draw a vertical line from the value 4 on the horizontal axis to the relevant concentration curve, and from that point read horizontally to the vertical axis coordinate. As can be seen, CR_4 is 40, 61.75, and 80, for A, B, and C, respectively. A similar exercise yields the eight-firm ratio CR_8 that also is often reported. Its value for markets A, B, and C is 80, 70.75, and 100, respectively.

Thus, for any n, we can define CR_n as the n-firm concentration ratio. A little thought will then quickly reveal that CR_n corresponds to a particular point on the industry's concentration curve. It follows that the principal drawback to such a measure is that it omits the other information in the curve. Compare for example the four-firm and eight-firm concentration

[1] Those familiar with the GINI Coefficient typically used to measure income inequality will recognize the concentration curve as the industrial structure analog of the Lorenz Curve from which the GINI Coefficient is derived. For further details, see Damgaard (2007).

ratios just given for the A, B, and C industries. Industry A appears more concentrated than does industry B using the CR_8 measure, but less concentrated when evaluated with the CR_4 index.

An alternative to CR_n that attempts to reflect more fully the information in the concentration curve is the Herfindahl-Hirschman Index or, more simply, the HI. For an industry with N firms, this is defined as follows:

$$\text{HI} = \sum_{i=1}^{N} s_i^2 \tag{3.1}$$

where s_i is the market share of the ith firm. In other words, the HI is the sum of the squares of the market shares of all of the firms in the industry. Table 3.1 illustrates the calculation of the HI for industry C in our example. If we measure market share in decimal terms so that a firm with 25 percent of the market has a share $s_i = 0.25$, the HI for industry C is 0.20. Compare this to a maximum value of HI = 1.0, which would be the HI if the industry were a pure monopoly with one firm accounting for all the output. However, the practice is often to measure the shares in percentage terms in which case the HI for industry C is 2000, which compares with a maximum, pure monopoly value of HI = 10,000 when shares are measured in this way. For industries A and B, similar calculations yield HI = 1,000 and HI = 3,126.25, respectively.

Like a concentration ratio, the HI measure has its drawbacks. However, it does have one very strong advantage over a measure such as CR_4 or CR_8. This is that the HI reflects the combined influence of both unequal firm sizes and the concentration of activity in a few large firms. That is, rather than just reflect a single point on the concentration curve, the HI provides, in a single number, a more complete sense of the shape of that curve. It is this ability to reflect both average firm size and inequality of size between firms that leads economists to prefer the HI to simple concentration ratios such as CR_4. In our example, Industry B gets the highest HI value because it is the one with the greatest disparity in firm sizes.

Table 3.1 Calculation of the HI for industry C

	Firm Rank	*Market Share (%)* s_i	*Squared Market Share* s_i^2
	1	25	625
	2	25	625
	3	25	625
	4	5	25
	5	5	25
	6	5	25
	7	5	25
	8	5	25
Sum:		100	2000 (HI)

3.1

Practice Problem

Consider two industries, each comprising ten firms. In industry A, the largest firm has a market share of 49 percent. The next three firms have market shares of 7 percent each, and the remaining six firms have equal shares of 5 percent each. In industry B, the top four firms share the bulk of the market with 19 percent apiece. The next largest firm accounts for 14 percent, and the smallest five firms equally split the remaining 10 percent of the industry.

a. Compute the four-firm concentration ratio and HI for each industry. Compare these measures across the two industries. Which industry do you think truly exhibits a more competitive structure? Which measure do you think gives a better indication of this? Explain.
b. Now let the three second largest firms in industry A merge their operations while holding onto their combined 21 percent market share. Recalculate the HI for industry A.

3.1.1 Measurement Matters: What Is a Market?

Whether one uses CR_4 or HI as an overall measure of a market's structure, it should be clear that the ability to make such measurements at all is predicated upon our ability to identify a well-defined market in the first place. In truth, this is not often easy to do. Consider, for example, the soft drink Pepsi. Does Pepsi compete only against other carbonated beverages, or should we also include such products as fruit juices, iced teas, and flavored milk in thinking about this market? Alternatively, think about multi-use products such as the *Xbox* that, among other capabilities, can function as a DVR and a platform for streaming services such as Netflix and HBO GO. Is the relevant market for *Xbox* simply that of video games, or does it also include products such as TiVo? Unless we have a clear procedure for answering such questions, any summary measure of market structure such as HI will become an arbitrary statistic capable of being manipulated either upward or downward at the whim of the researcher. An analyst can then make CR_4 or HI arbitrarily small or large by defining the market either broadly or narrowly.

One commonly used set of market definitions is the set maintained by the US Census Bureau (in collaboration with other agencies) known as the North American Industry Classification System (NAICS). The Bureau begins by first categorizing the output of business units in the United States into broad sectors of the economy, such as manufacturing, primary metals, agriculture, and forestry products, each of which receives a numeric code. These sectors are then subdivided further, and each is given a two-digit code. The manufacturing sector, for example, is covered by codes 31–33, while retail is covered by codes 44–45, and finance by 52. In each case, the two-digit codes are further disaggregated into the three-digit, four-digit, five-digit, and six-digit levels. Each additional digit represents a further subdivision of the initial classification. The Bureau surveys companies and assigns each plant to a specific six-digit industry, defined primarily on the basis of similarity of production processes. For plants that produce more than one product, the Bureau makes an assignment based on the plant's primary product, as measured by sales. Once all the establishments are so assigned at the most disaggregated level, it is easy to add up to the higher levels. It is also straightforward to work out the market shares and concentration indices according to who owns which plants for any industry at any level of aggregation.

Table 3.2 Concentration measures for selected industries

Industry	CR_4	*HI*
Breakfast Cereals	80.4	2425.5
Footwear Manufacturing	40.9	602.8
Automobiles	67.6	1448.8
Textile Mills	19.6	160.2
Paper Manufacturing	24.0	227.8
Petroleum Refineries	47.5	806.5
Pesticides	58.2	1109.1
Pharmaceuticals	29.5	359.1
Soap & Detergent Manufacturing	67.1	2025.3
Aluminum Sheet/Plate/Foil	70.5	1995.3
Computers & Peripherals	63.4	2030.7
Electric Light Bulbs & Parts	75.4	2258.3
Dolls, Toys, & Games	33.6	388.8
Aircraft	81.3	2652.2
Battery Manufacturing	58.1	1079.1
Telephone Apparatus	60.5	2078.5
Farm Machinery	59.0	1828.5
Tire Manufacturing	72.8	1539.6
Soft Drink Manufacturing	58.1	1094.5
Medical Equipment & Supplies	19.2	167.6

Source: "Concentration Ratios in Manufacturing," Bureau of the Census, 2007 Census of Manufacturing, January 2011.

These data are published regularly by the Census Bureau. Table 3.2 shows both CR_4 and HI for a sample of well-known NAICS industries.[2]

The two measures of industrial concentration, CR_4 and HI, are highly correlated, implying that each gives roughly the same description of an industry's structure. Yet while the CR_4 and HI measures often tell the same story, the crucial question is whether or not it is the right story.[3] As a little consideration quickly confirms, there is in fact reason to doubt that this is the case.

The source of the potential dissatisfaction with NAICS-based measures of industry structure is that these measures are based more on commonality of production technologies and less on whether the output of the different establishments considered really do compete with each other. Generally speaking, we would like to include production establishments in the same market if the goods that they produce are close substitutes in consumption. Yet as noted, the NAICS procedure is based more on a similarity of production techniques. For example, while wood flooring, resilient floor covering (e.g., linoleum tile), and ceramic tile are all flooring products, the NAICS categorization places them in different industries at the three-digit and all higher levels (321 for wood flooring, 326 for resilient flooring, and 327 for ceramic tile). In addition, the industries identified by the NAICS data calculate market shares on the basis of the national market when in fact practical considerations for a number

2 Further details are available online at http://www.census.gov/epcd/www/naics.html.

3 A quite readable discussion of the advantages and disadvantages of each ratio is available in Sleuwaegen and Dehandschutter (1986) and Sleuwaegen, Dehandschutter, and DeBondt (1989).

Reality Checkpoint

Concentrating on Concentration

Just as we can measure the fraction of an industry's output accounted for by its largest firms, so we can measure the fraction of the economy's entire output, GDP, accounted for by its largest corporations. However, while it may make some sense to speak of a concentration index based on just the top four to eight firms when speaking of a single industry, such a small number of firms would account for much too little of GDP to seriously consider. So, in the case of aggregate economic activity, we consider concentration ratios such as CR_{50} or CR_{200}. Such measures can be constructed using data from the Census Bureau's Census of Manufactures. Economist Lawrence White (2002) made such calculations for the US for various years up to the end of the 20th century. Some of his results are shown below.

Aggregate Concentration for Manufacturing (Value Added Basis); Selected Years, 1947–97

Year	CR_{50}	CR_{100}	CR_{200}
1947	17%	23%	30%
1958	23	30	38
1967	25	33	42
1977	24	33	44
1987	25	33	43
1992	24	32	42
1997	24	32	40

These data suggest that, at least since the 1950s, aggregate concentration in manufacturing has shown no increasing or decreasing trend. It was approximately the same in 1997 as it was in 1958 whether one looks at the top 50, 100, or 200 largest firms. White shows that somewhat similar results can be obtained if one looks at all nonfinancial corporations or focuses on shares of employment or profit. This of course does not mean that firms are not getting bigger. If each firm grows at the same rate as the economy, each of us will find ourselves employed in larger and larger organizations over time even though concentration is stable. White shows that this too has been happening and that the size of the average firm has, correspondingly, grown.

More recently, Elaine Tan (2008) has estimated both trends in concentration and trends in firm size inequality over a slightly longer time period than White (2002), namely, from 1931 to 2000, using total assets to measure firm size and assuming that the distribution of firm sizes is generated by a Champernowne distribution. Like its close cousin, the logistic distribution, the Champernowne distribution is like the normal distribution except that it has "fatter tails." Rather than look at the four-firm concentration ratio, Tan (2008) looks at the asset share of the largest 500 US firms both for all corporations and limited to non-financial corporations. In both cases, Tan finds that the share of total assets owned by the top 500 firms increased from 1931 to the middle of the second world war but then fluctuated with the end result that the share was about the same in 2000 (55 percent) as it was in 1942, although that share rose steadily during the 1990s. She also finds that the inequality among firm sizes has grown. Both of these results are consistent with White's earlier findings.

Sources: L. White. 2002. "Trends in Aggregate Concentration in the United States," *Journal of Economic Perspectives* (Fall): 137–60; and E. S. Tan. 2008. "Champernowne Model Estimates of Aggregate Concentration in the US, 1931–2000," *Working Paper*, University of London, Royal Holloway.

of goods, such as newspapers and ready-mix concrete, imply that inter-firm competition is only local or regional while for others such as passenger aircraft or automobile manufacture competition in international.

If true product markets are defined by the substitutability between the various goods sold, then one would ideally like to define such substitutes on the basis of a direct measure of such substitutability. A formal measure that captures this relationship is what economists call the cross-price elasticity of demand η_{ij}. This is defined as the percentage change in demand for good i that occurs when there is a 1 percent change in the price of another good j. The mathematical definition of this elasticity is

$$\eta_{ij} = \frac{\Delta q_i}{\Delta p_j} \frac{p_j}{q_i} \tag{3.2}$$

In brief, if η_{ij} is large and positive then goods i and j would be considered to be reasonably close substitutes.[4] Precisely what values of η_{ij} ought to be considered large and positive, though, is not immediately clear.

A practical approach to market definition that implicitly embodies the cross-elasticity measure is that employed by the Department of Justice and the Federal Trade Commission in evaluating mergers.[5] This is typically referred to as the SSNIP standard. This is an acronym for a "small but significant and non-transitory increase in price." Essentially, the authorities start with the narrowest possible definition of the market. For example, they might consider grocery stores selling natural, organic products such as Whole Foods. In that case, the SSNIP test would be whether, if all such natural foods grocery stores were monopolized, the monopolist could profitably raise the price of any of the goods by some small amount—usually 5 percent. If the answer is yes, then these grocery stores by themselves constitute a relevant product market. The intuition is that the hypothetical monopolist could only profitably raise prices if other stores left out of this market did not offer close substitutes and therefore the market analyzed is well-defined. If the answer to the SSNIP test is no, however, then the potential market is widened perhaps to all grocery stores natural or not, and the test of a profitable price increase by a hypothetical monopolist is repeated. If the answer is still no, then this process continues through ever wider definitions of the retail food market until the "yes" moment is realized.

While the SSNIP standard is clear however, its implementation is not. For example, in the early 2000s, Oracle, People-Soft, and SAP were the three leading firms engaged in developing, manufacturing, marketing, and servicing the complicated software used by businesses to manage their large organizations. This software is often further subdivided into separate modules for a corporation's many parts. Thus, there is a software module for human resources management (HRM), for financial management (FM), for supply chain management (SCM), and customer relations management (CRM). Besides the three firms already mentioned, there were also many smaller firms specializing in just one of these modules. The DOJ filed suit to block the proposed merger of Oracle and People-Soft. In its suit, the Justice Department argued that the relevant market based on an SSNIP test was the

4 However, a high monopoly price may inflate the cross-elasticity measure, a point originally emphasized by Stocking and Mueller (1955). Because the marginal consumer may look for substitutes when faced with a high monopoly price that would not be sought if the good were priced competitively, the cross price elasticity measured at the current industry price may be unduly large.

5 Since 1997, the European Commission has also explicitly recognized this standard as a tool for market definition.

market for high-end business software of the type sold to large and complex firms including major universities such as Stanford and Princeton, major banks such as Bank of America, and large government agencies such as the Department of Defense. The DOJ argued that competition to provide such complex firms with business software took place in auction markets in which only the three largest vendors could participate because each firm required a very detailed and uniquely configured software plan. The many smaller firms did not, and could not compete for these contracts because they focused on selling and servicing simple, standardized products that, while appropriate for much smaller firms, could never serve the needs of the large complex enterprises. The DOJ argued that an SSNIP test confirmed this market definition. However, Oracle disagreed and argued that a proper application of the SSNIP standard would indicate that a much broader market definition including all those firms preparing business software was appropriate.[6] Thus, while the standard may be clear, the proper way to impose it is more ambiguous.

3.1.2 More Measurement Matters: Geography and Vertical Relations

The application of the SSNIP standard above depended on what products define the market. Which products are included in the market and which are not? Clearly however, there is an equal need to define the market's geographic space. Consider for example the newspapers that serve the same city. Assuming that printed news media is determined to be a relevant product market, we still need to determine which printed media constitute the citywide newspaper market. Is it just newspapers printed and sold within the city? Or do national newspapers that are printed far beyond the city such as the *New York Times* and *USA Today* also constitute some of the market?

Alternatively, consider a region with a number of software producers who sell to a global market. If most of these firms' customers live outside the region then the geographic boundaries must be expanded. Similarly, it could be the case that local producers sell to a broad regional or even global market. As a result, although most local consumers buy from local producers, the presence of many consumers outside the local area means that the relevant geographic market extends beyond the local region.

In principle, an SSNIP test could be used to define the locations that constitute a geographic market just as it is used to define the goods that determine a specific product market. That is, once we have defined the relevant product market, we can define a region such as the local city and ask whether a firm that monopolized production within this area—say of newspapers or software—could impose small but significant and non-transitory increases in price. If the answer is yes, then the city does define a relevant geographic market. If the answer is no, then a wider geographic region needs to be considered.

In fact, the merger guidelines put out by the DOJ and the FTC point to exactly that kind of test. In practice however, alternative approaches are often used to define the relevant geographic market. A common test in this regard is the Elzinga-Hogarty (1978) test. This test considers two features of a geographic region to determine whether it represents a legitimate regional market. The first is LIFO (Little in from Outside) defined as 1 minus the fraction of all the products consumed within a region that are produced outside the region. Returning to our newspaper example, if 20 percent of all newspaper subscriptions within a local area are national newspapers, then LIFO would be 80 percent. The other measure of the Elzinga-Hogarty (1978) test is LOFI (Little out from Inside). This is calculated as

6 31 F. Supp. 2d 1098; 2004–2 Trade Cas. (CCH) ¶4,542. In the main, the court accepted Oracle's view and the merger was allowed to proceed.

1 minus the fraction of production within the region that is shipped to consumers outside the region. In other words, LIFO is a rough measure of 1 minus the fraction imported while LOFI is roughly 1 minus the fraction exported. Based on a review of historical evidence, Elzinga and Hogarty (1978) suggest that for regions in which either the LIFO and LOFI measures both exceed 75 percent, or where their average exceeds 90 percent, the presumption that the region is a well-defined geographic market is strong.

Yet while the Elzinga and Hogarty (1978) standard is easy to understand and measure, the threshold value needed to establish a definite market is not. As noted, Elzinga and Hogarty (1978) suggest a 75/90 threshold as a practical rule of thumb. Yet this critical value is based on economic intuition and a rough sense of what is workable—not on any probability distribution that would allow us to construct confidence intervals or test for statistical significance. It is not entirely clear, for example, why thresholds of 65 and 80 would not be better at capturing actual market boundaries. Note too that the definition of the product market has implications for the definition of the geographic market and vice versa. In the Oracle case, for example, the LIFO and LOFI measures vary depending on whether the product market consists of just the individualized business software sold by the big three firms or all business software sold by any firm.

A further issue with both product and market definitions relates to the relationships between firms operating at different stages of the production process. The delivery of a final good or service to the customer often represents the last of many steps. These include the acquisition of the raw materials, their transformation into a semi-finished good, the refinement of the semi-finished good into a final consumer product, and thereafter, the retailing. In economics jargon, the initial raw materials phase is typically described as the upstream phase after which the product flows "downstream" through the various stages toward its final sale to the consumer. The relationship between upstream and downstream phases is therefore a vertical one, and there are several forms that this relationship can take. An upstream producer may own all the subsequent phases in which case we say the firm is vertically integrated. Alternatively, an upstream producer may offer franchising agreements or long-term contracts to downstream sellers. The existence and variability of such relationships can cause difficulty in measuring the structure of the market at any one stage of production. For instance, there are many bottling companies so that conventional measures of market concentration in the bottled can and soft drink industry are rather low. In turn, this suggests a fairly competitive market. However, the reality is that most bottling companies do not compete with each other but, instead, are tied through strict contractual agreements to use one of the national upstream suppliers, such as Coca-Cola or Pepsi.[7] As a result, there may be much less competition among bottlers than the concentration measure would suggest.

In sum, the interpretation of structural market measures such as CR_4 and HI is greatly complicated by a variety of factors beginning with the fact that any such structure is endogenous. Much of our data collection organizes markets on the basis of similar production techniques rather than on a measure of substitution. Implementing more appropriate market measures such as SSNIP requires careful application and interpretation of the evidence, and this can be problematical. Geographic definitions such as the Elzinga and Hogarty (1978) standard are easy to understand, but it is not obvious which values of that standard should serve as a threshold of market determination. The point is that while much of the theoretical work covered in this text will take the market at hand to be well understood, the real world measurement of actual markets is fraught with difficulties.

7 Some authors, for example, Gort (1962) and, more recently, Davies and Morris (1995), have tried to obtain a precise, quantitative measure of the extent of vertical integration.

At the same time, one should not overstate the case. For example, categorizing industries on the basis of closeness of shared production techniques does not necessarily indicate the substitutability necessary for competition. However, it has the advantage that it likely does group firms with similar production costs. This is important because any analysis that links an industry's configuration to the underlying structure of its technology and costs only makes sense if the production technologies are sufficiently similar that we can make general, industry-wide statements about a typical firm's cost structure.

3.2 MEASURING MARKET POWER

Throughout this chapter, we have been thinking about market structure in the quite literal sense of how the industry's production of output is allocated across different firms. We have seen how summary statistics such as the CR_4 or HI attempt to describe this configuration of firms in an industry much as a census taker might use similar statistics to describe the number and size of families in a geographic region. A large part of the motivation for these measures is the desire to summarize succinctly just where an industry might lie relative to the ideal of perfect competition. There is nothing wrong with this structural approach so long as one clear caveat is kept in mind: that a particular structure does not necessarily imply a particular outcome.

When we say that an industry is highly concentrated we are saying that the industry output is dominated by a few firms, in contrast to the configuration that we associate with the competitive model. Does that necessarily mean then that prices charged in this industry are above what would prevail in a perfectly competitive market? The answer is not so straightforward. As we shall see in subsequent chapters, markets with even just two or three firms may sometimes come quite close to duplicating the competitive or efficient outcome.

The Lerner Index is one way to measure how well a market performs from an efficiency point of view. The Lerner Index, *LI*, measures how far the outcome is from the competitive ideal in the following way:

$$LI = \frac{P - MC}{P} \tag{3.3}$$

Because the Lerner Index directly reflects the discrepancy between price and marginal cost it captures much of what we are interested when it comes to the *exercise* of market power. For a competitive firm, the Lerner Index is zero because such a firm prices at marginal cost. For a pure monopolist, on the other hand, the Lerner Index can be shown to be the inverse of the elasticity of demand—the less elastic the demand the greater the price-marginal cost distortion. (See the Appendix of Chapter 2 for a formal derivation.) To see this, recall that for a monopolist the marginal revenue of selling an additional unit of output can be written as $MR = P + \frac{\Delta P}{\Delta Q}Q$. For profit maximization we set marginal revenue equal to marginal cost, or $P + \frac{\Delta P}{\Delta Q}Q = MC$. Rearranging and dividing by price P we obtain

$$\frac{P - MC}{P} = -\frac{\Delta P}{\Delta Q}\frac{Q}{P} = \frac{1}{\eta} \tag{3.4}$$

where $1/\eta$ is the inverse of the elasticity of demand. The less elastic is demand, or the smaller is η, the greater is the difference between market price and marginal cost of production in the monopoly outcome. To drive the point home, recall that the perfectly competitive firm faces an infinitely elastic or horizontal demand curve. When such a large value is substituted for the elasticity term, equation (3.4) implies a Lerner Index of 0. Again, the perfectly competitive firm sells at a price equal to marginal cost. Note too that the Lerner Index can never exceed 1 and that it can only hit this maximum value if marginal cost is 0.

For an industry of more than one but not a large number of firms, measuring the Lerner Index is more complicated and requires obtaining some average index of marginal cost. A particularly straightforward case in this regard is that in which the commodity in question is homogenous so that all firms must sell at exactly the same price. If this is so, then we can measure a market-wide Lerner Index as:

$$LI = \frac{P - \sum_{i=1}^{N} s_i MC_i}{P} \tag{3.5}$$

Here, as before, s_i is the market share of the ith firm and N is the total number of firms.

The Lerner Index is a very useful conceptual tool and we will make reference to it throughout the remainder of this book. Like the CR_4 or the HI, the Lerner Index is a summary measure. The difference is that the Lerner Index is not so much a measure of how an industry's production is structured as it is a measure of the market outcome. The greater the Lerner Index, the farther the market outcome lies from the competitive case—and the more market power is being exploited. In this sense, the Lerner Index is a direct and useful gauge of the extent of market competition.

For example, Ellison (1994) tries to get evidence on game-theoretic models of cartel behavior. For this purpose, he studies railroad prices over time in the late nineteenth century. He estimates that, apart from price war periods, the Lerner Index was about 85 percent of what it would be under pure monopoly pricing. In other words, the collusive behavior of railroads at this time was capable of coming within 15 percent of the pure monopoly price distortion.

However, much like the structural indices, the Lerner Index also has its problems. To begin with, calculating the Lerner Index for an industry runs into the problem of market definition. Further, even when the market definition is reasonably clear, the Lerner Index is still difficult to measure. It is one thing to count the number and estimate the sizes of the various firms in an industry. Measuring the elasticity of demand is trickier. Measuring marginal cost is even more difficult. Unfortunately, even small changes in the assumptions one makes about the data can lead to sizable differences in estimated price–cost margins. For example, Ellison's (1994) study relied on data studied earlier by Porter (1983). Yet Porter's (1983) estimate of the price distortion during collusive periods is only half as large as Ellison's (1994) estimate.

Moreover, even when the Lerner Index is accurately measured its interpretation can remain ambiguous. Suppose for example that each firm in an industry has to incur a one-time sunk cost F associated with setting up its establishment. Assume further that each firm's marginal cost is constant. Because each firm needs to earn enough operating profit to cover its sunk cost, the equilibrium price level will need to rise above marginal cost. That is, the Lerner Index will need to be positive. However, the more positive that difference is—the greater is the price–cost margin—the greater the number of firms that can cover the one time sunk cost. As a result, we might observe a high Lerner Index in a setting in which there are numerous firms, none of which is very large. In such a case, the high Lerner

Index might erroneously indicate little competition even though no firm has any significant market power.[8]

Conversely, the Lerner Index might underestimate market power in settings in which cost-reducing innovations are important. Suppose for example that an industry has an old and not very efficient incumbent firm with high marginal cost. As long as demand is somewhat elastic, such a firm may have no choice but to price relatively close to marginal cost. At the same time, the incumbent has a great incentive to take whatever actions it can that will keep a low-cost rival from entering the market. In this case, the Lerner Index deceptively indicates a fair bit of competition because price is low relative to the incumbent's marginal cost when the relevant but unavailable comparison is the price with the potential rival's lower marginal cost.[9]

3.3 EMPIRICAL APPLICATION: MONOPOLY POWER—HOW BAD IS IT?

A recurrent question in antitrust policy is just how costly imperfect competition is for the economy overall. If the losses from monopoly power are not large, then devoting any significant resources to antitrust enforcement to prevent such losses is probably not worthwhile. Such scarce resources would be better used in, say, increasing homeland security or providing relief to hurricane victims. If the economic costs of market power are large, however, then allocating resources to combat the abuse of that power is likely to be warranted. Hence, it would be useful if economists had some sense of just how serious the losses from monopoly power actually are.

In principle, economists have a clear measure of the economic loss caused by monopoly power. It is the deadweight loss or triangle that results from prices above marginal cost. In practice, however, measuring this loss is not so easy. This is because it requires getting estimates of cost and/or demand but, as with any estimate, these values are subject to some error. Unfortunately, rather small changes in the estimates can lead to rather large changes in the estimated welfare cost.

To understand the issues involved, let us start with the basic measurement of the welfare or deadweight loss that results from pricing above marginal cost. As shown in Chapter 2, this is the area whose height is given by the difference between price P and marginal cost MC, and whose base is given by the difference between the competitive output Q^C that would sell if $P = MC$ and the actual market output Q that sells at the actual price P. Hence, the welfare loss WL is:

$$WL = \frac{1}{2}(P - MC)(Q^C - Q) \tag{3.6}$$

It is convenient to express this welfare loss as a proportion of total sales revenue PQ to yield

$$WL' = \frac{WL}{PQ} = \frac{1}{2}\frac{(P - MC)}{P}\frac{(Q^C - Q)}{Q} \tag{3.7}$$

8 See, for example, Elzinga (1989).
9 Hovenkamp (1994), among others, has made this argument.

Remember that the elasticity of demand η is the proportionate increase in output in response to a given proportionate decrease in price. If the price were to fall from its current P level to the competitive level of $P = MC$, then output would rise to the competitive level of Q^C. That is:

$$\eta = \frac{(Q^C - Q)/Q}{(P - MC)/P} \quad \Rightarrow \quad \frac{(Q^C - Q)}{Q} = \eta \frac{(P - MC)}{P} \tag{3.8}$$

Because we also know that the industry Lerner Index is $(P - MC)/P$, we can rewrite equation (3.7) as:

$$WL' = \frac{WL}{PQ} = \frac{1}{2}\eta (LI)^2 \tag{3.9}$$

Now recall from equation (3.4) earlier in the chapter that, for a pure monopolist, the Lerner Index is given by: $LI = (P - MC)/P = 1/\eta$. Then, in this case, the deadweight loss relative to industry sales will be:

$$WL' = \frac{WL}{PQ} = \frac{1}{2}\frac{1}{\eta} \tag{3.10}$$

That is, for the perfect monopoly case, the deadweight loss as a fraction of current industry sales is simply one-half the Lerner Index or one over twice the elasticity of demand. The intuition is that as the demand elasticity increases, the welfare loss shrinks because other goods are increasingly viewed as substitutes to the monopolized commodity. Note further the sensitivity of the welfare loss to the elasticity estimate. An estimate that $\eta = 1.5$ produces a welfare loss equal to 33 percent of revenue. An estimate of $\eta = 2$ reduces this amount to 25 percent of revenue. That is, a 0.5 change in the elasticity estimate yields an eight percent change in the welfare loss.

The first person to make calculations along the foregoing lines on a large scale was Arnold Harberger (1954). Using a sample of 73 manufacturing industries, Harberger (1954) took the difference in the five-year average industry rate of return and the five-year average for manufacturing overall as an approximation of LI. Because he worked with industry data, and because none of the industries was a pure monopoly, Harberger (1954) could not assume that his LI estimate is the inverse of elasticity of demand, as we did in equation (3.10). Instead, he combined his LI estimates with an assumed demand elasticity of $\eta = 1$ in equation (3.9). The dollar value of these estimated distortions is then given by multiplying WL' by industry sales PQ. When Harberger (1954) added these dollar values up and extrapolated the results across the entire economy he found a surprisingly small welfare cost of monopoly—on the order of one-tenth of one percent of Gross Domestic Product. The low value of Harberger's (1954) estimate thus raised a serious question about the cost-effectiveness of antitrust policy and litigation.

Harberger's (1954) approach, however, did not go uncriticized. Bergson (1973) noted that Harberger's (1954) procedure essentially used a partial equilibrium framework to obtain a general equilibrium measure. He demonstrated that, in principle, this could mean that Harberger's (1954) estimate considerably understated the actual loss. Cowling and Mueller (1978) used firm-level data for 734 companies in the United States and 103 companies in the United Kingdom. The use of firm-level data means that Cowling and Mueller (1978) could apply equation (3.10) directly. Their estimated monopoly welfare costs range from 4 to 13 percent of GDP in the United States and from 4 to 7 percent in the United Kingdom—far larger than Harberger's (1954) estimates.

Table 3.3 Hall's (1988) estimated Lerner Index for selected industries

Industry	*Lerner Index*
Food & Kindred Products	0.811
Tobacco	0.638
Textile Mill Products	−0.214
Apparel	0.444
Lumber and Wood	0.494
Furniture and Fixtures	0.731
Paper and Allied Products	0.930
Printing	0.950
Rubber & Plastic	0.337
Leather Products	0.524
Stone, Clay, and Glass	0.606
Primary Metals	0.540
Fabricated Metals	0.394
Machinery	0.300
Electric Equipment	0.676
Instruments	0.284
Miscellaneous Mfg	0.777
Communication	0.972
Electric, Gas & Sanitary Services	0.921
Motor Vehicles	0.433
Average	0.57

Robert Hall (1988) used a production theory approach to derive estimates of the Lerner Index for twenty broad manufacturing sectors in the United States. These are shown in Table 3.3. Domowitz, Hubbard, and Petersen (1988) obtained similar but generally lower estimates of the Index using Hall's (1988) approach corrected for changes in raw material usage. Whereas Hall (1988) found an average price–cost margin of 0.577, Domowitz, Hubbard, and Petersen (1988) estimate the average to be only 0.37. Even this lower value, however, indicates a substantial degree of welfare loss in the manufacturing sector due to non-competitive pricing.

An important source of variation in Cowling and Mueller's (1978) analysis is how advertising costs are treated in measuring LI. This calls attention to the critical importance of the marginal cost measure in determining welfare losses. This issue has been addressed more recently by Aiginger and Pfaffermayr (1997). They start by recognizing that without the pressure of perfect competition, firms can operate in an industry with different cost efficiencies. Thus, the average industry marginal cost $\overline{MC}$ is very likely not the minimum average cost that would be enforced if perfect competition were the rule. Aiginger and Pfaffermayr (1997) then make use of a result (one that we shall derive in Chapter 9) from a standard oligopoly model. The result is that the industry price-cost margin measure using $\overline{MC}$ is equal to the industry Herfindahl Index HI (scaled from 0 to 1), divided by the elasticity of industry demand. That is:

$$\frac{P - \overline{MC}}{P} = \frac{HI}{\eta} \quad \Rightarrow \quad \eta = HI\left(\frac{P}{P - \overline{MC}}\right) \tag{3.11}$$

Substituting this result into equation (3.9), we obtain:

$$WL' = \frac{WL}{PQ} = \frac{1}{2}\left(\frac{P - MC}{P}\right)\left(\frac{P - MC}{P - \overline{MC}}\right) HI \tag{3.12}$$

Note that the term $\left(\dfrac{P - MC}{P - \overline{MC}}\right)$ is greater than one because MC is the marginal cost that would prevail under competition. Aiginger and Pfaffermayr (1997) measure this competitive MC as the marginal cost of the most efficient firm in the industry under the assumption that this is the cost efficiency that would be required for competitive firms to survive. Effectively, their approach permits them to decompose the welfare cost of market power into two parts. One is the traditional welfare loss measure due to prices not equal to *industry* average marginal cost, $P - \overline{MC}$. The other is due to the fact that market power allows the survival of firms with higher than minimum costs, $\overline{MC} - MC$. Using data from 10,000 cement and paper firms in the European Union, Aiginger and Pfaffermayr (1997) find that the total welfare loss of market power in these industries is on the order of 9 to 11 percent of industry sales. Perhaps not surprisingly, they find that these welfare losses are largely due to the cost inefficiencies that imperfect competition permits. Thus, their estimate of the traditional welfare loss measure is on the order of 2 to 3 percent, while the cost inefficiency loss is on the order of 7 to 7.5 percent. Extrapolating these estimates to the entire economy would yield results that are considerably closer to the Cowling and Mueller estimates (1978) than those obtained by Harberger (1954).

In evaluating all of these estimates, it is useful to bear in mind at least two caveats. First, an implicit assumption in all these calculations is that it is feasible to have perfect competition in all industries. As we shall see in the next chapter, however, costs and technology make this an unlikely outcome. In this sense, the estimates of welfare losses due to monopoly price distortions are too high as there is no way in which all industries could be freed of such market power. Second, the measures are taken from data in which active antitrust enforcement has been the norm. In this sense, the measures are an understatement of the potential for monopoly-induced welfare losses. Had there been no antitrust enforcement, there would have presumably been more market power abuses and the associated welfare losses would have been greater.

Summary

This chapter has focused on the measurement of market structure and market power. We are very often interested in summarizing the extent to which an industry departs from the competitive ideal in a single number or index. The issue then becomes whether and how we can construct such a summary measure.

Concentration indices, such as the CR_4 or HI, are explicit measures of a market's structure. Both look at firm shares as a fraction of the industry's total output. Both encounter important problems, such as the difficulty of accurately defining the relevant market. The HI, however, is generally preferred by economists since it not only reflects the top firm shares but also the differences in relative firm sizes.

An explicit measure of market power is the Lerner Index. Because it is based on a comparison of price and marginal cost, this index directly addresses the extent to which the market outcome deviates from the competitive ideal. However, the need to measure marginal cost accurately, along with other measurement issues, makes the Lerner Index as difficult to employ as the structural indices. Estimates of the Lerner Index also serve as a useful starting point to estimate the actual efficiency costs of monopoly power. Many efforts have been made to do this for the entire economy in an attempt to get a general view as to how serious the problem of market power really is. These empirical studies have yielded a wide range of estimates of the aggregate deadweight loss as a percentage of GDP. The lower bound estimate is that monopoly power imposes only a small inefficiency cost of a few tenths of one percent of GDP. However, upper bound estimates range as high as 14 percent. A crucial parameter in such studies is the elasticity of demand assumed to be typical.

As long as the foregoing problems are recognized, the CR_4, HI, and Lerner Index measures are useful starting points to characterize an industry's competitive position. However, an industry's degree of concentration and price–cost margin do not materialize out of thin air. Instead, these indices all derive from the interaction of a number of factors. One of those factors is the nature of production costs. The role that technology and cost play in shaping the industrial outcome is examined in the next chapter.

Problems

1. The following table gives US market share data in percentages for three paper product markets in 1994.

Facial Tissue		*Toilet Paper*		*Paper Towels*	
Company	% Share	Company	% Share	Company	% Share
Kimberly-Clark	48	Proctor & Gamble	30	Proctor & Gamble	37
Proctor & Gamble	30	Scott	20	Scott	18
Scott	7	James River	16	James River	12
Georgia Pacific	6	Georgia Pacific	12	Georgia Pacific	11
Other	9	Kimberly-Clark	5	Scott	4
		Other	16	Other	18

   a. Calculate the four-firm concentration ratio for each industry.
   b. Calculate each industry's HI Index.
   c. Which industry do you think exhibits the most concentration?

2. Consider a market comprised of three firms. Firm 1 produces and sells 23 units per period. Firm produces and sells 19 units per period, while firm 3's periodic production and sales are 15 units. The (inverse) market demand is estimated to be: $P = 100 - Q$, where $Q =$ total output $= q_1 + q_2 + q_3$. Determine the market price and the elasticity of market demand as well as the market share of each firm. Now use equation (3.5) to determine each firm's marginal cost. What relation do you find between marginal cost and market share?

3. Monopoly Air is on record with the local transportation authority arguing that there is no market need and no market room for an additional air service because even now, Monopoly Air planes are flying with only 60 percent of the seats typically booked. Hence, it argues that the market is not large enough to sustain two efficient-sized air carriers.

   Evaluate this argument. Why might Monopoly Air flights be so under-booked? Does this prove that there is no "market room" for a new competitor?

4. We defined the Lerner Index as $LI = 1/\mu$ where μ is the absolute value of the elasticity of demand. We also showed that LI can be alternatively expressed as $(P - MC)/P$. Use these relationships to show that LI can never exceed 1. What does this imply is the minimum demand elasticity we should ever observe for a monopolist?

References

Aiginger and Pfaffermayr. 1997. "Looking at the Cost Side of Monopoly," *Journal of Industrial Economics* 44 (September): 245–67.

Bergson, A. 1973. "On Monopoly Welfare Losses," *American Economic Review* 63 (December): 853–70.

Cowling, K., and D. C. Mueller. 1978. "The Social Cost of Monopoly Power," *Economic Journal* 88 (December): 727–48.

Damgaard, C. 2007. "Lorenz Curve." From *MathWorld*–A Wolfram Web Resource, created by Eric W. Weisstein. http://mathworld.wolfram.com/LorenzCurve.html

Davies, S. W., and C. Morris. 1995. "A New Index of Vertical Integration: Some Estimates for UK Manufacturing," *International Journal of Industrial Organization* 13:151–78.

Domowitz, I., R. G. Hubbard, and B. Petersen. 1988. "Market Structure and Cyclical Fluctuations in Manufacturing," *Review of Economics and Statistics* 70 (February): 55–66.

Ellison, G. 1994. "Theories of Cartel Stability and the Joint Executive Committee," *Rand Journal of Economics* 25 (Spring): 37–57.

Elzinga, K. 1989. "Unmasking Monopoly Power: Four Types of Economic Evidence," in R. Larner and J. Meehan, Jr., eds., *Economics and Antitrust Policy*. Westport, CT: Greenwood Press.

Elzinga, K., and Hogarty. 1978. "The Problem of Geographic Market Delineation Revisited: The Case of Coal," *Antitrust Bulletin* 23 (Spring): 1–18.

Gort, M. 1962. *Diversification and Integration in American Industry*. Princeton: Princeton University Press.

Hall, R. 1988. "The Relation Between Price and Marginal Cost in U. S. Industry," *Journal of Political Economy* 96 (October): 921–47.

Harberger, A. 1954. "Monopoly and Resource Allocation," *American Economic Review* 45 (May): 77–87.

Hovenkamp, H. J. 1994. *Federal Antirust Law Policy: The Law of Competition and Its Practice*. St. Paul: West Publishing.

Porter, R. 1983. "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886," *Bell Journal of Economics* 14 (Autumn): 301–14.

Sleuwaegen, L., and W. V. Dehandschutter. 1986. "The Critical Choice between the Concentration Ratio and the H-Index in Assessing Industry Performance," *Journal of Industrial Economics* 35 (December):193–208.

Sleuwaegen, L., W. V. Dehandschutter, and R. DeBondt. 1989. "The Herfindahl Index and Concentration Ratios Revisited," *Antitrust Bulletin* 34 (Fall): 625–40.

Stocking, G., and W. Mueller. 1955. "The Cellophane Case and the New Competition," *American Economic Review* 45 (March): 29–63.