Chapter 8
Market Structure, Industry Concentration, and Barriers to Entry

In economics, we normally classify markets into four market structures: perfect competition, monopoly, monopolistic competition, and oligopoly. In this chapter, we are interested in understanding why real markets are structured so differently. For example, most agricultural commodities approximate competitive markets, as they have many producers of homogeneous or nearly homogeneous goods. In contrast, the market of computer operating systems is nearly monopolized by Microsoft. In 2009, Microsoft Windows had a market share of approximately 92%, while its nearest competitor, Mac, had a market share of just over 5%.

We will see that in many cases market structure is relatively stable over time, although this is not always the case. Technological change can transform industry structure by giving large scale producers a cost advantage and put smaller competitors out of business. This is what happened in the US brewing industry, where the number of traditional brewers declined from 476 in 1945 to about 19 today. The internationalization of the automobile industry led to more foreign cars being sold in the USA. Entry caused the market share of the dominant domestic car companies (General Motors or GM, Ford, and Chrysler) to fall from over 90% in the mid 1950s to approximately 55% in the late 2000s. Thus, we are also interested in understanding how market forces cause market structure to change over time.

How a market is structured can have important welfare implications. We learned in Chaps. 5 and 6 that perfectly competitive markets are allocatively efficient, while monopoly markets are allocatively inefficient. This suggests that static inefficiency rises with less competition, a viewpoint that is consistent with the structure–conduct–performance paradigm discussed in Chap. 1. It is also consistent with the tenor of the Sherman and Clayton Acts, which are designed to support competitive market structures. Although we will see that the hypothesis that an increase in the number of competitors improves welfare is not always correct, understanding the reasons why a market has just a few competitors and why their numbers may change over time will give us a better understanding of the nature of competition in a dynamic world.

V.J. Tremblay and C.H. Tremblay, *New Perspectives on Industrial Organization*, Springer Texts in Business and Economics, DOI 10.1007/978-1-4614-3241-8_8,

8.1 The Delineation of Market Structure

Before discussing the qualities of market structure more generally, we first review the characteristics of the four market structures found in microeconomics textbooks (see Table 8.1). To begin with, profit maximization is assumed throughout. The extreme cases of perfect competition and monopoly were discussed in Chaps. 5 and 6. Recall that in perfect competition there are many producers and goods are homogeneous. Entry and exit are free, and firms are price takers, meaning that the market price is exogenously determined. In a monopoly, entry barriers make it possible for only one firm to enter the market. In this case, the firm is a price maker, meaning that the firm has the power to raise price without losing all of its customers. Although most markets lie between these polar cases, competitive and monopoly models provide us with useful reference points, that is, extremes in market structure that identify lower and upper bounds on the expected equilibrium price in a market.

As we discussed in Chap. 6, monopolistic competition has characteristics of both monopoly and perfect competition. Like perfect competition, entry is free and there are many competitors in the market. A key feature of monopolistic competition is the presence of product differentiation, which gives each firm a monopoly over the sale of its particular brand. Thus, we can think of monopolistic competition as a competitive market with product differentiation or a monopoly market with free entry of closely related goods.

The market structure that has received little attention so far is **oligopoly**. In an oligopoly market, products may or may not be differentiated, and entry barriers are present. The key feature of oligopoly is that only a few firms account for the bulk of industry production. Because strategic interaction is important, with one firm's actions affecting its own profits and the profits of its competitors, game theory is used to develop oligopoly models. The steel and aluminum markets are examples of oligopoly markets with homogeneous goods. The automobile and cell phone industries are examples of differentiated oligopolies.

When asked which of the four market structures are most common in the USA, many students choose monopolistic competition. This response is understandable because most consumer goods markets have many differentiated brands. But typically only a handful of firms produce most brands in a particular market.

Table 8.1 Characteristics of the four main market structures

Market structure	Number of firms	Product type	Entry/exit	Price
Perfect competition	Many	Homogeneous	Free	Exogenous
Monopoly	1	Just 1 Product	B.E.	Endogenous
Monopolistic competition	Many	P.D.	Free	Endogenous
Oligopoly	Few	Homogeneous & P.D.	B.E.	Endogenous

Note: B.E. refers to barriers to entry, and P.D. refers to product differentiation. Exogenous implies that firms have no control over price; endogenous implies that firms have a least some control over price.

These are called multibrand or multiproduct producers, as opposed to the single product producers discussed in most elementary textbooks.

A classic example is the market for breakfast cereal. In most supermarkets you can find over a hundred brands of cereal. Brands are made from a variety of grains (e.g., oats, corn, wheat, bran, and rice), can come in a variety of flavors (e.g., brown sugar, cinnamon, honey, chocolate, strawberry, and peanut butter), and may contain raisins, dried strawberries, dried peaches, or nuts. Yet, the largest five cereal companies produce most of these brands, accounting for 94% of cereal sales in 2008. Although industries such as these appear to be monopolistically competitive, they are actually oligopolies, the most common market structure.

Given that oligopolies are so common, we devote most of our attention to understanding these markets. In this chapter, we begin with a discussion of the concept of industry concentration, which characterizes the extent to which industry production is concentrated in the hands of a few firms in an industry. In Sect. 8.3 we discuss the extent of industry concentration in the USA. In Sect. 8.4 we investigate the main determinants of industry concentration. That is, why do some markets have many firms and others just a few firms? We will also summarize the empirical evidence regarding the causes of high industry concentration. In later chapters, we investigate how equilibrium price and output are determined in oligopoly markets and compare these outcomes with those found in competitive and monopoly markets.

8.2 Industry Concentration

In this section, we summarize the principle methods of measuring industry concentration and discuss their strengths and weaknesses. Because proper measurement requires that a market be correctly defined, we also discuss the issue of market definition.

8.2.1 The Meaning and Measurement of Industry Concentration

A prominent feature of market structure is industry concentration. The number and size distribution of firms within an individual market indicates the extent of concentration.

One way to visualize industry concentration is with a **concentration curve**. A concentration curve plots the cumulative market share of sales that are attributable to the largest through the smallest firms in the industry.[1] To illustrate, consider three hypothetical industries (A, B, and C), which have six, eight, and ten competitors, respectively. Output and market share information for each industry

[1] The market share for a particular firm is defined as the firm's sales divided by industry sales, where sales are typically measured by output or by total revenue.

Table 8.2 Firm output, market share, and industry concentration for three hypothetical industries: A, B, and C

	Output (1,000,000 s)			Market share (%)			Squared market share		
Firm	A	B	C	A	B	C	A	B	C
1	4.8	4.8	3.6	40	20	10	1,600	400	100
2	2.4	4.8	3.6	20	20	10	400	400	100
3	1.2	4.8	3.6	10	20	10	100	400	100
4	1.2	4.8	3.6	10	20	10	100	400	100
5	1.2	1.2	3.6	10	5	10	100	25	100
6	1.2	1.2	3.6	10	5	10	100	25	100
7		1.2	3.6		5	10		25	100
8		1.2	3.6		5	10		25	100
9			3.6			10			100
10			3.6			10			100
Industry	12	24	36						

	Industry A	Industry B	Industry C
n	6	8	10
CR_4	80	80	40
HHI (MS_i measured in %)	2,400	1,700	1,000
HHI′ (MS_i measured as a decimal)	0.24	0.17	0.10
Numbers Equivalent ($n' = 1/HHI'$)	4.17	5.88	10.0

Note: MS_i is firm i's market share, which can be measured in percent or as a decimal.

is listed in Table 8.2. The concentration curves for these industries are plotted in Fig. 8.1. They reveal two important facts. First, a concentration curve is a straight line when each firm is of equal size, as in industry C. Second, the curve shifts up with fewer competitors and as larger firms gain market share. For instance, industries A and B have fewer firms and higher concentration curves than industry C. In addition, the concentration curve for industry A starts at a higher point than for industry B (and C), because the market share of the largest firm is 40% in industry A and 20% in industry B (10% in industry C). Thus, we can conclude that higher industry concentration is reflected in a higher concentration curve.

Although a concentration curve provides a clear picture of concentration, economists have also tried to create a single index of industry concentration. A single index is useful for empirical work and for addressing antitrust concerns. Ideally, a concentration index should take into account the size distribution of all firms in the industry. It should also increase, implying greater concentration, when the number of firms declines and when a larger firm gains market share from a smaller firm.

Given this criteria, the **number of firms** (n) in an industry is an unsatisfactory index of concentration unless all firms within an industry are of equal size. When this is not the case, two industries with 100 firms would be considered equally concentrated even if one industry had firms of equal size and the other had a large firm with a market share of 95%. A firm such as this is called a **dominant firm** because it typically takes a leadership role in choosing price or output due to its

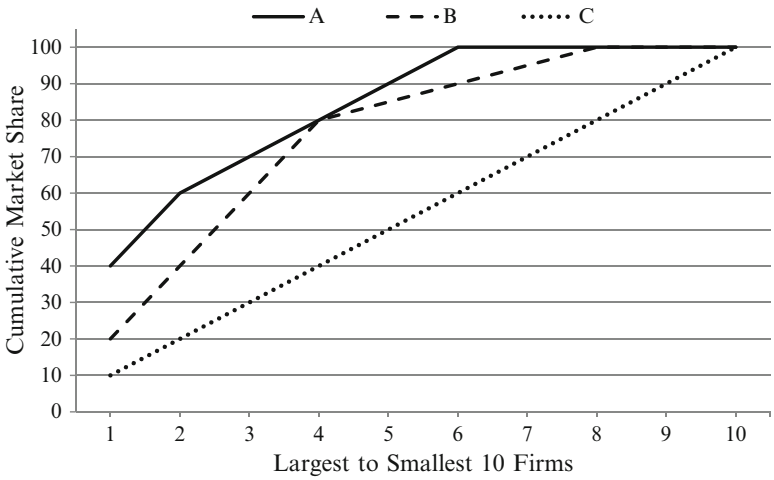

Fig. 8.1 Concentration curves for hypothetical industries A, B, and C

large market share relative to its competitors, which are sometimes called competitive fringe firms. Dominance can result from producing a superior product or producing at lower cost than its competitors. Because the dominant firm in this example has a near monopoly, we would like our index to reflect a higher level of concentration in this case than in the symmetric case where firms are of equal size.

A more commonly used index is the ***k*-firm concentration ratio** (CR_k), defined as the market share of the k largest firms in the industry. If we order firms from the largest (firm 1) to the smallest (firm n), the k-firm concentration ratio is

$$CR_k \equiv \sum_{i=1}^{k} ms_i, \tag{8.1}$$

where ms_i is firm i's market share (which can be measured as a decimal or percent), ms_1 is the market share of the largest firm, ms_2 is the market share of the second largest firm, etc., and ms_k is the market share of the kth largest firm. Notice that CR_k approaches 0 as the number of equal sized firms increases and approaches 1 or 100% when the k largest firms supply more and more of the industry's output. The four-firm concentration ratio (CR_4) has been used by economists for decades because it is regularly calculated for a variety of industries by the US Census of Manufacturers. Notice that a concentration ratio is represented by a point on a concentration curve. In Fig. 8.1, CR_4 is found by identifying the cumulative market share on the vertical axis associated with firm four on the horizontal axis. For industries A and B it is 80%, and for industry C it is 40%.

The main advantage of a concentration ratio is that it is easy to calculate and understand. Unfortunately, it suffers from three main weaknesses. First, it provides no information about the relative shares of the largest k firms. Second, it completely ignores the distribution of sales outside the largest k firms. As an example, a merger

of firms 7 and 8 in industry B from Table 8.2 will have no effect on CR_4 even though the number of firms has diminished and the distribution of output among firms has changed (i.e., the merged firm becomes the fifth largest firm with a market share of 10%). Third, concentration ratios do not always provide consistent rankings of industry concentration. One can see from Fig. 8.1, for example, where industries A and B are equally concentrated if we use CR_4, but industry A is more concentrated than industry B if we use CR_1, CR_2, CR_3, CR_5, CR_6, or CR_7.

An alternative index of industry concentration is the Herfindahl–Hirschman Index (HHI).[2] Mathematically,

$$\text{HHI} \equiv \sum_{i=1}^{n} \text{ms}_i^2. \tag{8.2}$$

When market share is expressed in percent, then HHI approaches 0 in a competitive market and equals 10,000 for a monopoly. To illustrate, based on the squared market share figures in Table 8.2, HHI equals 2,400 for industry A, 1,700 for industry B, and 1,000 for industry C. When market share is expressed as a decimal, HHI ranges from 0 to 1. In this case, HHI equals 0.24 for industry A, 0.17 for industry B, and 0.1 for industry C.

Unlike a concentration ratio, the HHI meets our criteria for a desirable concentration index. In particular, it decreases with the number of firms (n) and increases with the variance in market share (σ^2). When we *measure market share in decimal form*, we can rewrite (8.2) as[3]

$$\text{HHI} = n\sigma^2 + 1/n. \tag{8.3}$$

This demonstrates that HHI increases as the variance in market share increases.

Equation (8.3) also implies that when firms are of equal size, so that $\sigma^2 = 0$, then HHI $= 1/n$. HHI has a value of 1 in a monopoly market and diminishes as n increases and firms remain equal in size. Equation (8.3) can give us a **numbers equivalent**, such that a given value of HHI can be translated into a number of equal sized firms (n'). When market share is measured as a decimal, the numbers equivalent of a given value of HHI is $n' = 1/\text{HHI}$. Values of n' are calculated in Table 8.2 for our hypothetical industries. For industry A, n' is 4.17 (i.e., 1/.24),[4] which means that for HHI to equal 0.24, there would need to be 4.17 equal sized firms in the industry. This provides another way of thinking about HHI. Because of

[2] For a discussion of the history of this index, see Hirschman (1964).

[3] To see this, note that the variance (σ^2) can be written as $\sigma^2 = [\Sigma ms_i^2/n - (\Sigma ms_i/n)^2]$; market shares sum to 1 (when measured in decimals), so that $\Sigma ms_i = 1$; HHI $= \Sigma ms_i^2$. Thus, $\sigma^2 = \text{HHI}/n - 1/n^2$. Solving for HHI gives HHI $= n\sigma^2 + 1/n$. For a discussion of variance, see the Mathematics and Econometrics Appendix at the end of the book.

[4] When market share is measured in decimals, note that HHI is 0.24 or 2,400/10,000. The numbers equivalent is 1/0.24 or 4.17. These are frequented rounded off to the nearest counting number, which would be 4 in this case.

Table 8.3 Industry concentration and market classification in a market with n' equal size firms

					Market classification	
n'	$CR_1 = 100 \cdot 1/n'$	CR_4	HHI	HHI′	SRS	Merger guidelines
1	100	100	10,000	1.00	T–O	H
2	50.0	100	5,000	0.50	T–O	H
3	33.3	100	3,333	0.33	T–O	H
4	25.0	100	2,500	0.25	T–O	H
5	20.0	80.0	2,000	0.20	T–O	H
5.56	18.0	72.0	**1,800**	0.18	T–O	**H**
6	16.7	66.7	1,667	0.67	T–O	M
6.67	15.0	**60.0**	1,500	0.15	**T–O**	M
7	14.3	57.1	1,429	0.43	O	M
8	12.5	50.0	1,250	0.13	O	M
9	11.1	44.0	1,111	0.11	O	M
10	10.0	**40.0**	**1,000**	0.10	**O**	**M**
11	9.1	36.4	909	0.09	C	Un
20	5.0	20.0	500	0.05	C	Un
100	1.0	4.0	100	0.01	C	Un

Note: Market share is measured in percent for the one-firm concentration ratio (CR_1), the four-firm concentration ratio (CR_4), and the Herfindahl–Hirschman index (HHI). When market share is measured in decimal form: HHI′ = HHI/10,000, $n' = 1/\text{HHI}'$, CR_1 and CR_4 must be divided by 100, and the figures for HHI above must be divided by 10,000.

Regarding market classification. SRS refers to the Scherer and Ross (1990) and Shepherd (1997, 16) market classifications: tight oligopoly (T–O) when CR_4 reaches 60%, oligopoly (O) when CR_4 reaches 40% and is less than 60%, and competitive (C) when CR_4 is less than 40%. These cutoffs are in bold in columns 3 and 6.

Merger Guidelines refers to the Department of Justice and the Federal Trade Commission classification: a market is unconcentrated when HHI is less than 1,000 (Un), moderately concentrated when HHI ranges from 1,000 to less than 1,800 (M), and highly concentrated when HHI is greater than or equal to 1,800 (H). These cutoffs are in bold in columns 4 and 7.

these desirable features, the Department of Justice began using HHI as a measure of industry concentration in 1982 to evaluate potential antitrust violations.[5]

Concentration ratios and HHI are the most commonly used indices of industry concentration, and it is useful to investigate their properties further. When *market share is measured in percent* and all firms are of equal size,

$$\begin{aligned} CR_4 &= \min(100, 400/n) \\ \text{HHI} &= 10,000/n. \end{aligned} \tag{8.4}$$

Furthermore, when $n \leq 4$, $\text{HHI} = (100 \cdot CR_4)/n$; when $n > 4$, $\text{HHI} = 25 \cdot CR_4$. If we *measure market share as a decimal*, then $\text{HHI} = CR_1 = 1/n$. To make this more concrete, in Table 8.3 we list several examples for an industry where firms are of equal size and market share is measured in percent.

Experts in the field have identified critical values of concentration indices that distinguish competitive from oligopoly markets. Scherer and Ross (1990, 82) and

[5] The obvious drawback with HHI is that it requires sales data on every firm in the industry.

Shepherd (1997, 16) contend that once CR_4 reaches 40%, strategic interaction becomes significant and an industry can be classified as an oligopoly. Once CR_4 reaches 60%, Shepherd classifies it as a tight oligopoly, one where collusion is likely. When enforcing the antimerger laws, the Department of Justice and the Federal Trade Commission use the following delineation:

- An industry is classified as unconcentrated when HHI is less than 1,000.
- An industry is classified as moderately concentrated when HHI ranges from 1,000 to less than 1,800.
- An industry is classified as highly concentrated when HHI is greater than or equal to 1,800.[6]

As Table 8.3 indicates for firms of equal size, CR_4 equals 40% when HHI equals 1,000, and CR_4 equals 72% when HHI equals 1,800. Thus, there is some consistency among experts.

8.2.2 Definition of the Relevant Market

When measuring concentration, a crucial step is to properly define the market. In fact, many antitrust decisions hinge on how broadly or narrowly a market is defined. If defined too broadly, firms will be included that are not true competitors and our concentration index will be biased downwards.

A relevant economic market includes all products that are close substitutes in consumption and production. Defining a market requires that we draw proper product and geographic boundaries. Geographically, markets may be local, regional, national, or international. Typically, this depends on the value of the product, its weight, and shipping costs per mile. A product will ship a longer distance as its unit shipping cost falls and as its value to weight ratio increases.

Several examples illustrate this idea. Diamonds are shipped worldwide, while cement is rarely shipped more than 150 miles. There are thousands of cement suppliers nationally, but only a few firms are true suppliers in any particular region in the country. Thus, if we incorrectly define the cement market as national, our estimate of industry concentration would be biased downwards. In contrast, the automobile market is international in scope, with domestic producers GM, Ford, and Chrysler accounting for about half of US automobile sales. If we ignore foreign competitors, then our estimate of industry concentration will be too high because it will ignore imported cars from Japan, Germany, and other countries.

Correctly defining the product boundary is equally important. If all products were reasonably homogeneous and distinct, product boundaries would be relatively clear: a banana supplier competes with other banana suppliers, and a peanut butter

[6] US Department of Justice and the Federal Trade Commission, *Horizontal Merger Guidelines*, April 2, 1992 and April 8, 1997. A comparison of the old with the new structural standard is difficult, because the new 2010 Merger Guidelines have more lenient standards and consider a broader set of factors. For further discussion of the 2010 Guidelines, see Chap. 20.

supplier competes with other peanut butter suppliers. Product boundaries become fuzzy, however, when products are imperfect substitutes.

When discussing product differentiation in Chap. 7, we said that the market includes goods that perform the same basic function, even though there are slight differences among brands (i.e., they have slightly different characteristics). Clearly, different brands of men's athletic shoes should be included in a market, but what about men's shoes and women's shoes? For most consumers, men's and women's shoes are poor substitutes. Another example is salt, where most suppliers produce a homogeneous good. Yet, road salt is not a substitute for table salt.

One way to identify a group of closely substitutable products is to estimate the cross-price elasticity of demand between products. Recall from Chap. 2 that the cross-price elasticity of demand between products i and j (η_{ij}) is defined as the percentage change in the quantity demanded of product i with respect to a small change in the price of good j. More formally, it is given by

$$\eta_{ij} \equiv \frac{\partial q_i}{\partial p_j} \frac{p_j}{q_i}. \tag{8.5}$$

The value of η_{ij} tells us how sharply demand for good i changes in response to an increase in the price of good j. When η_{ij} is large and positive, products i and j are considered close substitutes. We would anticipate a sizable cross-price elasticity for Coke and Pepsi, but what about Coke and Mountain Dew or Coke and orange juice? We would expect that the cross-price elasticity will be higher as we compare Coke to other brands of cola versus all brands of soft drinks or all beverages. Even with accurate estimates of η_{ij}, there is no clear cutoff value that we can use to decide which products belong to a particular market. Some judgment is required.

A more practical approach may be to consider the price movements of a class of like products in a particular geographic region. If prices are similar and move together over time, then products within the class are more likely to be close substitutes. For example, a 2010 Honda Civic Coupe is of similar size to a 2010 Porsche 911, but their price difference ($18,000 versus $79,000) indicates that they are in different markets. If one were to ask Honda dealers who are Civic competitors, they would likely identify a Ford Focus (retailing at $16,000), Subaru Impreza ($17,500), Toyota Corolla ($17,000), and VW Jetta ($18,000), not a 911 Porsche.

When investigating possible antitrust violations regarding horizontal mergers, the Department of Justice and the Federal Trade Commission have their own approach to defining a market.[7] According to their guidelines, a product's competitors include:

- All products to which buyers would switch if a firm raised the price of its product by 5%.
- The products of all potential competitors that would be expected to enter the market within 1 year if all existing firms raised their prices by 5%.

[7] US Department of Justice and the Federal Trade Commission, *Horizontal Merger Guidelines*, April 2, 1992 and April 8, 1997.

Although somewhat speculative, this definition acknowledges the importance of potential competition. A year is a short time though, so under this criterion potential competitors would include only those firms that can easily transform existing production capacity from one market to another. For example, it may be relatively quick and easy for a table salt producer to make the conversion to road salt production than for a road salt producer to make the conversion to table salt production.

To summarize, a relevant economic market should include all products that are close substitutes in consumption and production. Delineating a market requires that we draw appropriate geographic and product boundaries and consider all potential entrants.

8.3 The Extent of Industry Concentration in US Markets

In this section we discuss the degree of industry concentration in the USA. We begin by reviewing trends in **aggregate concentration**, the market share of total US sales that are produced by the largest corporations. Next, we list CR_4 and HHI for a sample of well-known industries. Finally, we analyze the trend in concentration for a single industry, the US brewing industry. The results show that aggregate concentration has been relatively stable over time; the level of concentration differs across industries; and concentration can change dramatically over time for an individual industry.

8.3.1 Aggregate Concentration

The leading US corporations have grown to enormous size, and their flagship brands are internationally recognized. In the USA, Wal-Mart was the largest in 2007, with total revenue of $378 billion. Of the top 5, three are oil companies: Exxon Mobil (number 2), Chevron (3), and ConocoPhillips (5). The fourth is GM. Although some of the largest corporations focus on a single market, most are conglomerates, and their size does not necessarily translate to high concentration in any one industry. Nevertheless, there are concerns that large corporate size generates considerable economic power and political clout.

In spite of this concern, the evidence indicates that aggregate concentration has been fairly constant since the late 1950s. White (2002) investigated this issue by compiling data on the total market share (in terms of value added) for the largest 50 (CR_{50}), 100 (CR_{100}), and 200 (CR_{200}) corporations in the manufacturing sector of the economy. Although aggregate concentration rose between 1947 and 1958, the three measures were remarkably stable from 1958 to 1997 (see Table 8.4). The reader should be aware, however, that the largest US corporations have grown in absolute size as the overall economy has expanded.

Table 8.4 Aggregate concentration of the largest 50, 100, and 200 corporations in the manufacturing sector of the US economy

Year	CR_{50}	CR_{100}	CR_{200}
1947	17	23	30
1958	23	30	38
1963	25	33	41
1967	25	33	42
1970	24	33	43
1977	24	33	44
1982	24	33	43
1987	25	33	43
1992	24	32	42
1997	24	32	40
Mean	23.5	31.5	40.6

Note: Concentration for the manufacturing sector is based on value added (total revenue minus the cost of materials) for the largest 50, 100, and 200 corporations in the USA.
Source: White (2002).

Table 8.5 An example of NAICS subcategories

NAICS Code	Subdivision	Description
31	Sector	Manufacturing
312	Subsector	Beverage and tobacco manufacturing
3121	Industry Group	Beverage manufacturing
312111	Industry	Soft drink manufacturing
312112	Industry	Bottled water manufacturing
312120	Industry	Beer manufacturing
312130	Industry	Wine manufacturing
312140	Industry	Distilled spirits manufacturing

Note: Six-digit codes are used outside the USA.
Source: US Census Bureau, "Concentration Ratios in Manufacturing: 2002," at http://www.census.gov/prod/ec02/ec0231sr1.pdf

8.3.2 Concentration for Selected Industries

One source of concentration data is the US Bureau of the Census. The Census Bureau periodically publishes CR_4 and HHI, based on the North American Industry Classification System (NAICS).[8] The manufacturing and services areas of the economy are split into 20 sectors and are identified by two-digit codes. These are subdivided further into 100 subsectors (identified by three-digit codes), 317 industry groups (four-digit codes), and 1,179 industries (six-digit codes). Table 8.5 provides an example of the NAICS subdivisions for food manufacturing for various

[8] This system of classifying industries has been in effect since 1997. Prior to 1997, data were published according to the Standard Industrial Classification (SIC) system.

beverage industries. In terms of product boundary, the six-digit code comes closest to what we would call an economic market, such as soft drink manufacturing (312111) and wine manufacturing (312130).

Table 8.6 lists values of CR_4 and HHI for a set of well-known industries. The data show that concentration varies widely from industry to industry. Concentration for textile mills is very low, while concentration is extremely high in the market for electric light bulbs. The data also reveal a high degree of correlation between these two indices of concentration. In this sample, the correlation coefficient is 97.0%. In other studies for different samples and time periods, the correlation coefficient between CR_4 and HHI ranges from 0.929 to 0.992.[9]

The main drawback with the Census estimates of industry concentration is that they are based on the assumption that markets are national in scope. This geographic boundary is frequently incorrect, however. As discussed above, the market for cement is local, not national; therefore, the true level of industry concentration is higher than those found in Census estimates. Alternatively, the automobile industry in international in scope, and the national Census measures of concentration are too high. As a result, Census estimates of industry concentration must be used with caution.

8.3.3 Changes in Concentration for a Single Industry

Early studies following the structure–conduct–performance tradition maintained that market structure was exogenous and relatively stable over time (Bain 1956, 1959). As Table 8.7 reveals, CR_4 was relatively stable from 1963 through 1997 for petroleum refineries, pharmaceuticals, cement, tires and tubes, and soap and other detergents.[10] Nonetheless, critics of the structure–conduct–performance paradigm contend that market structure can be endogenous and change substantially over time.[11]

The US brewing industry has witnessed extensive changes in concentration since the 1930s. After the end of Prohibition in 1933, the number of independent mass-producing beer companies reached a peak at just over 700 brewers in 1938. These include companies such as Anheuser-Busch, Miller, Coors, and Pabst that brew traditional American lager beer, such as Budweiser, Miller Lite, Coors Light, and Pabst Blue Ribbon. Since then, the number has steadily declined to about 19 independent brewers today.[12]

---

[9] For a review of these studies, see Kwoka (1985), Scherer (1980, 58–59), and Scherer and Ross (1990, 72–73).

[10] A detailed comparison for all industries listed in Table 8.6 is not possible because some industries are defined differently in the NAICS system and the older SIC system.

[11] In particular, see Demsetz (1973), Peltzman (1975), and Sutton (1991). For a review of the literature, see Scherer (1980, Chap. 9) and Martin (2002, Chap. 6).

[12] This excludes microbrewers or specialty brewers that make European style ales and lagers and began entering the market in the mid 1960s. Although the number of specialty brewers exceeds 1,600 today, their combined market share is less than 6% and they generally compete for a different type of customer.

Table 8.6 Concentration indices for selected industries

Industry	CR_4	HHI
Textile mills	13.8	94
Sporting and athletic goods	21.4	161
Plastic pipes/fittings	24.8	241
Frozen fruit, juice, and vegetables	34.3	350
Book printing	32.0	364
Meat products	35.0	393
Petroleum refineries	28.5	422
Ice cream and frozen desserts	32.3	445
Iron and steel mills	32.7	445
Pharmaceutical and medicine	32.3	446
Computer and peripheral equipment	37.0	465
Cement	33.5	467
Dolls, toys, and games	40.0	496
Toiletries	38.6	564
Cookies, crackers, and pasta	41.7	602
Computers	40.0	658
Semiconductors	41.7	689
Women's footwear (except athletic)	49.5	795
Soft drinks	47.2	800
Men's and boy's suits and coats	42.0	846
Men's footwear (except athletic)	49.7	857
Telephone equipment	55.3	1,061
Distilleries	60.0	1,076
Aluminum sheet/plate/foil	65.0	1,447
Tires and inner tubes	68.4	1,518
Soap and other detergents	65.6	1,619
Household refrigerators and freezers	81.5	2,025
Automobiles	79.5	2,350
Breakfast cereals	82.9	2,446
Aircraft	84.8	–
Electric light bulbs	88.9	2,849
Motor vehicles and car bodies	87.0	–
Breweries	89.7	–
Cigarettes	98.9	–

Note: CR_4 is the four-firm concentration ratio measured in percent, and HHI is the Herfindahl–Hirschman index.
Source: US Census Bureau, "Concentration Ratios in Manufacturing: 2002," at http://www.census.gov/prod/ec02/ec0231sr1.pdf

Table 8.7 A sample of industries for which the four-firm concentration ratio (CR_4) is stable over time

	CR_4	
Industry	1963	1997
Petroleum refineries	34	29
Pharmaceuticals	22	32
Cement	29	34
Tires and inner tubes	70	68
Soap and other detergent	72	66

Source: US Census Bureau, "Concentration Ratios in Manufacturing," at http://www.census.gov/prod/ec02/ec0231sr1.pdf

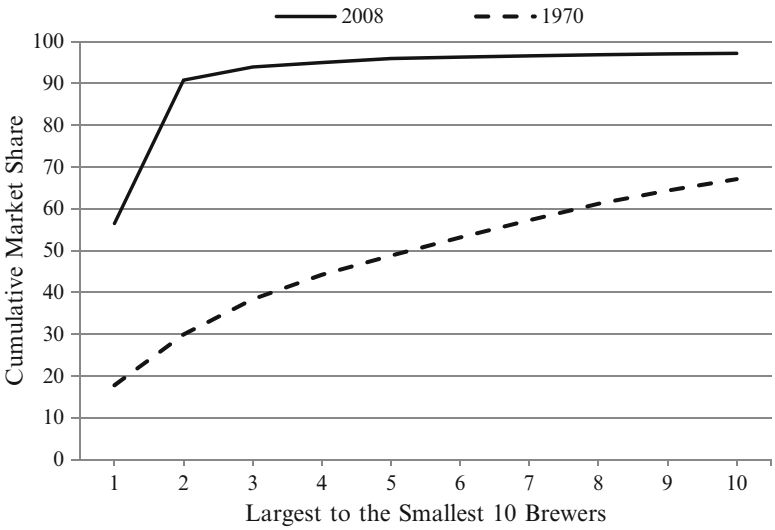

Fig. 8.2 Concentration curves for largest ten brewing companies

One way to see that concentration has increased in brewing is to compare concentration curves over time. Figure 8.2 plots concentration curves for the largest ten brewing companies in 1970 and 2008. Recall that a concentration curve identifies the cumulative market share, in this case based on total domestic beer consumption. Notice that the concentration curve is substantially higher in 2008, reflecting an increased level of industry concentration. It is also more convex in 2008 than in 1970 due to the fact that the largest firms now controlled a much larger share of the market. For example, the cumulative market share of the largest 2 firms (CR_2) was 30% in 1970 and 91% in 2008.[13]

The pattern of rising concentration in brewing can also be seen in Fig. 8.3, which plots CR_4 and HHI from 1947 to 2008. Both series reveal a dramatic and almost continuous increase in concentration.[14] For example, CR_4 rose from 44 to 94% and HHI rose from 7.08 to 43.29 from 1970 to 2008. Consistent with studies using data from other industries, the correlation coefficient between CR_4 and HHI is quite high, at 0.962.

The data in Figs. 8.2 and 8.3 must be interpreted with caution before 1970. CR_4 and HHI are for the nation as a whole, but the market was regional in scope until the

[13] In 1970, the two largest firms were the Anheuser-Busch and Miller brewing companies. In 2008, they were Anheuser-Busch and MillerCoors (the combined sales of the Miller and Coors brewing companies which formed a joint venture in 2008).

[14] To compare it to CR_4, HHI is divided by 100 so that it ranges from 0 to 100.

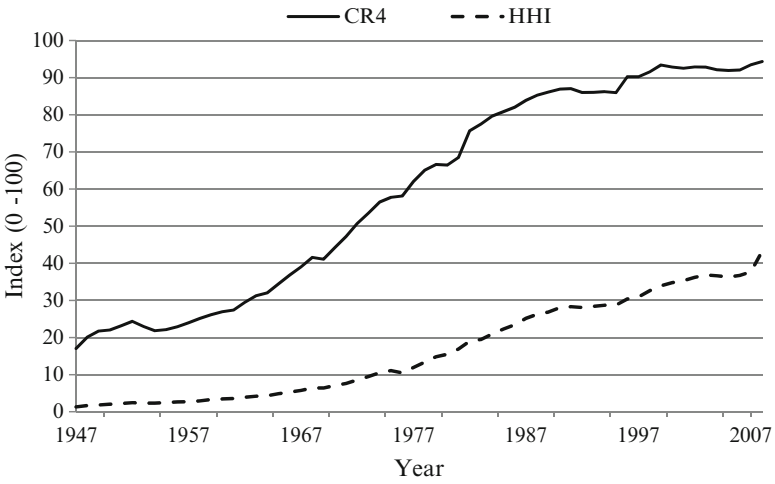

Fig. 8.3 The four-firm concentration ratio and the Herfindahl–Hirschman Index in the US brewing industry, 1947–2008

late 1960s.[15] Given that the market was national by 1970, the Merger Guidelines would classify the brewing industry as unconcentrated before 1974, moderately concentrated from 1974 through 1981 (when HHI rose from 0.1053 to 0.1691), and highly concentrated from 1982 on (with HHI exceeding 0.1800 after 1981).

8.4 The Determinants of Market Structure

We have seen that the level of industry concentration varies across industries and can change considerably over time. In this section, we investigate the main reasons why concentration is high in some industries and low in others.

8.4.1 Gibrat's Law

One of the simplest reasons why industry concentration may increase over time was proposed by Gibrat (1931), who considered the effect of luck on concentration.[16] His analysis assumes an industry that initially had a fixed number of equal sized firms. For our purposes, assume 50 firms, each with a market share of 2%, implying a CR_4 of 8% and an HHI of 200. Over time, firms experience an increase in sales and face the same growth distribution. In this example, the distribution is normal, with a mean

[15] For a more complete discussion of the geographic market in brewing, see V. Tremblay and C. Tremblay (2005, Chap. 3).

[16] For an excellent review of the influence of Gibrat's work, see Sutton (1997).

growth rate of 6% and a standard deviation of 16%.[17] This means that firm growth is simply a random event. The key point is that even though the average growth rate is 6%, some firms will be lucky and grow at a faster rate than average, while others will be unlucky and grow at a slower rate. Given these circumstances, Gibrat asked whether or not industry concentration would remain constant over time.

The answer is somewhat surprising. As time goes on, some firms gain market share due to a string of good luck, while persistently unlucky firms lose share. Thus, even though firms start out the same and face the same distribution in growth rates, firm size becomes skewed over time, approaching a lognormal distribution.[18] As a result, the concentration curve shifts upwards, implying higher concentration. This is called Gibrat's Law of Proportionate Effect, or simply **Gibrat's Law**. To illustrate Gibrat's Law, Scherer and Ross (1990, 141–146) ran a simulation of an industry with these characteristics. They found that the distribution of firm size became more skewed over time and that CR_4 rose from period to period, starting out at 8% in period 1 and averaging 54.7% by period 140.

One concern with Gibrat's Law is that it provides no economic rationale for industry concentration; it is simply due to pure chance. There are certainly aspects of business where luck is important. As we will see in Chap. 14, there is an element of luck with advertising. Ex ante, all firms have high expectations for their upcoming advertising campaigns even though only a fraction of them are successful ex post. Thus, lucky firms with successful ad campaigns experience greater growth than their competitors. Nevertheless, advertising agencies would argue that there is more to successful advertising than pure luck.

Consequently, it is highly unlikely that Gibrat's Law is the only explanation for high concentration. After all, industry concentration does not always rise over time. We have already seen in Table 8.7 that concentration has remained relatively constant for petroleum refineries, pharmaceuticals, cement, tires and tubes, and soap and other detergents. In addition, concentration has fallen in some industries. Thus, other forces must also come into play. One example is greater globalization, which can decrease concentration by increasing the number of competitors and reducing the market share of industry leaders, as in the automobile industry. A second example is technological change, which can lower concentration if it favors smaller firms or raise concentration if it favors larger firms.

Another mark against Gibrat's Law is that the prediction that the size distribution will be lognormal does not appear to be true empirically. After reviewing the evidence, Schmalensee (1989, 994) concludes that "all families of distributions so far fail to describe at least some industries well." Thus, the process generating firm size distributions appears to be more complex than that postulated by Gibrat.

Perhaps the strongest piece of evidence to suggest that chance is not the only determinant of concentration is a fairly consistent pattern of industry concentration

[17] See the Mathematics and Econometrics Appendix at the end of the book for a review of a normal distribution and a standard deviation.

[18] In a lognormal distribution, the logarithm of firm size is normally distributed.

Table 8.8 The four-firm concentration ratio by country

	Country						
Industry	USA	France	Germany	Italy	Japan	UK	Mean
Processed Meat[a]	19	23	22	11	51	–	25
Bread[a]	25	5	7	4	48	58	25
Sugar confectionary	27	51	39	29	48	38	39
Sugar[a]	46	81	69	72	42	94	67
Canned vegetables[a]	50	40	–	80	–	81	63
Flour[a]	55	29	38	7	67	78	46
Pet food	64	86	93	–	39	83	73
Biscuits	68	62	49	46	49	62	56
Mineral water	–	77	27	55	62	73	59
Soup	75	91	84	–	71	75	79
Beer	81	82	25	55	99	59	67
Salt[a]	82	98	93	80	–	99	90
Breakfast cereal	86	–	–	–	–	79	83
Soft drinks	89	70	57	84	88	48	72
Baby foods	90	88	83	88	–	80	86

Source: Sutton (1991, 106).

[a]Sutton identifies these markets as having relatively homogeneous goods and receiving little advertising support. The remaining are classified as advertising-intensive industries.

across nations (Schmalensee 1989, 992). In their study of six nations,[19] Scherer et al. (1975) found that the markets for cigarettes, bottles, refrigerators, and batteries tended to be highly concentrated in every nation, while the markets for weaving, paints, and shoes tended to be unconcentrated in every nation. In a more recent study, Sutton (1991) finds very similar results (see Table 8.8). This evidence indicates that when industry concentration is high (low) in one nation, it tends to be high (low) in others.

Although luck may be a factor, previous evidence is sufficiently strong to conclude that systematic forces play a dominant role in shaping industry concentration. For this reason, we focus the remainder of our attention on market and strategic rather than random forces that can influence market structure.

8.4.2 Concentration and Barriers to Entry

Fundamentally, entry conditions play a key role in determining industry concentration. Perfectly competitive markets have many producers because the cost of entry and exit is zero. In contrast, barriers to entry insulate a sole firm in a monopoly market from competition. In this section, we discuss in more detail what is meant by a barrier to entry and outline the primary types of barriers that restrict entry and lead to high levels of industry concentration.

[19] These are the USA, Canada, the UK, Sweden, France, and Germany.

Economists have defined the concept of a barrier to entry in several different ways.[20] Bain (1959) defines it as a market condition that raises the cost of entering the market to such an extent that incumbent firms earn long-run economic profits. Of course, even a monopoly firm can lose money in the short run and earn zero profit in the long run. Stigler (1968) argues that a barrier to entry exists only if the cost of entry is higher for new entrants than it was for established firms. Finally, von Weizsacker (1980) defines a barrier as a limitation on entry that is socially undesirable.

On the surface, one might think that any constraint on entry is socially undesirable, but this need not be the case. For instance, a patent gives a firm a 20 year monopoly to a new invention, thus eliminating all entry. Yet, this barrier to entry is generally thought to be socially beneficial, because it encourages innovation and dynamic efficiency. Although von Weizsacker's welfare based definition of a barrier to entry is appealing, its main weakness is that it substantially complicates our use of the concept. His perspective does remind us though that if we define a barrier as a cost of entry, then we are ignoring its welfare implication (Martin 2002, 343).

In this book, we take a pragmatic approach, defining a **barrier to entry** to include any limitation on entry that keeps the long-run equilibrium number of firms below the number that would exist in a competitive market. With this definition, there are no barriers to entry in the perfectly competitive and monopolistically competitive models because they both have many competitors. Barriers do exist in a monopoly market with just one firm and in an oligopoly market with just a few competitors. Again, this definition is consistent with what most people mean by entry barrier and is easy to use, but it does not rule out the possibility that a particular barrier is welfare enhancing.

Baumol et al. (1982) show that entry barriers are closely linked to sunk costs, expenditures that cannot be recovered if the firm exits the market. Suppose that you plan to start a new business that requires a $1 billion investment. You apply for a loan, and the first thing that the loan manager asks is what you will put up for collateral. Unless you are extremely wealthy, your answer will depend on your plans for the money. If you are purchasing a factory that will be worth $1 billion if you were to go out of business, then there is no sunk cost associated with the investment and you can use the factory as collateral. If, on the other hand, your investment is speculative, such as hiring scientists to find a cure for the common cold, then most if not all of your investment is a sunk cost. If unsuccessful, the money invested evaporates and is not recoverable. Of these two investment opportunities, which do think would be easier to raise the $1 billion? Obviously, the investment with no sunk cost carries no risk to you or the bank, and financing would be relatively easy to obtain. Accordingly, entry barriers are closely tied to sunk costs.

Another important aspect of entry barriers is that they can be either exogenously or endogenously determined. By exogenous barriers we mean that firms in the industry

[20] For examples of different definitions of barriers to entry, see Bain (1956), Stigler (1968), von Weizsacker (1980), and McAfee et al. (2004).

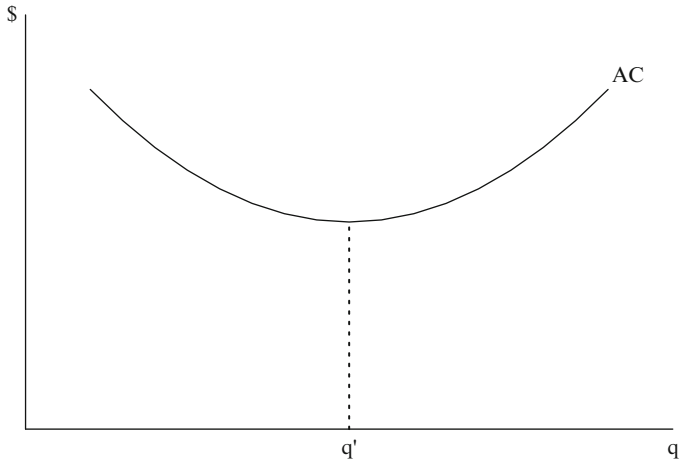

Fig. 8.4 Long-run average cost curve and minimum efficient scale (MES)

have no control over them. Exogenous determinants of entry barriers might derive from basic demand and cost conditions, because demand conditions are determined by consumers and cost conditions are technologically determined.[21] They can also include government regulations that legally restrict entry. Examples include a patent or a government franchise that limits the number of competitors, such as your local cable television company. Barriers that are caused by basic economic conditions are called **natural barriers to entry**. Those that are caused by government restrictions are called **legal barriers to entry**. We postpone discussion of them until Chap. 20.

Barriers that are endogenous are sometimes called **strategic barriers to entry** because they are under the control of firms in the industry and are specifically designed to deter entry. These include a variety of predatory activities that are profitable only because they drive existing competitors out of business or deter potential competitors from entering the market. Examples include predatory pricing, where price is cut below unit cost, and actions that raise rival costs. In the sections below, we illustrate how natural and strategic entry barriers affect industry concentration.

8.4.2.1 Concentration and Natural Barriers to Entry

Natural barriers are determined by market demand and cost conditions. We saw in Chap. 2 that economies of scale exist when the long-run average cost (AC) curve has a negative slope. This is illustrated in Fig. 8.4, where there are economies of scale until output reaches q'. Beyond q', AC has a positive slope, indicating

[21] Of course, firms could invest in research and development, which can change technology and lead to an increase or a decrease in scale economies.

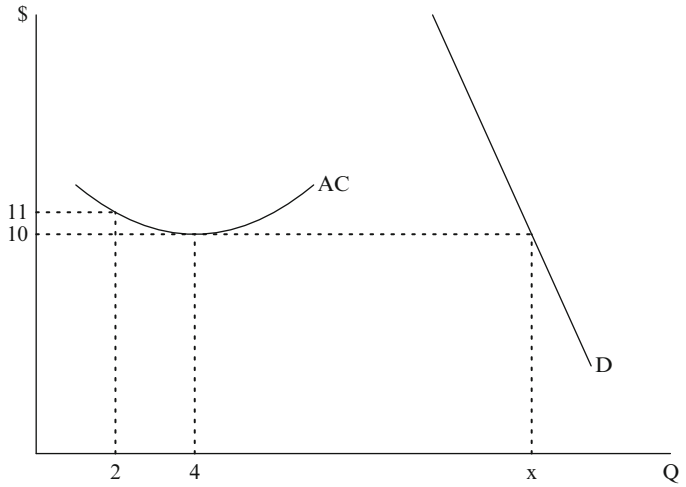

Fig. 8.5 Demand and cost conditions that support a natural oligopoly

diseconomies of scale. Recall from Chap. 2 that the smallest output for which AC is at its minimum, q', is called minimum efficient scale (MES).

One way to see how demand and cost conditions affect entry barriers and concentration is to review the theory of a natural monopoly (discussed in Chap. 6). A natural monopoly occurs when there are substantial scale economics relative to the size of the market (represented by market demand), making it productively inefficient to have more than one firm produce total market output. If there are many firms, each firm can lower its cost by merging with a competitor, a process that will continue to be profitable until there is just one firm left in the market. In this case, demand and cost conditions make it productively efficient and most profitable for a single firm to serve the market.

We can generalize this idea to the case of n firms by considering the demand and cost structure described in Fig. 8.5. Consistent with the notation used previously, AC is long-run average cost and D represents market demand. In this example, MES equals 4 (million units), which corresponds to an average cost of $10. Baumol et al. (1982) define the **cost-minimizing industry structure** as the number of firms in an industry that are needed to produce industry output (x) at minimum cost, which equals x/MES $= n^*$.[22] When this occurs, the industry is productively efficient. To demonstrate, when x equals 20, five symmetric firms minimize the total cost of producing $x = 20$ by each producing at MES $= 4$. Thus, the cost-minimizing industry structure is five firms.

In this example, notice that total industry cost is not minimized when the number of firms differs from five. Take the case of ten symmetric firms, each producing two units. In this case, AC $= \$11$, and the total industry cost of producing 20 million

[22] Here, we assume that x/MES produces an integer, thus avoiding problems with fractions.

units is $220 million. With five firms, AC = $10, and the total industry cost of producing 20 is only $200 million. Thus, firms have an incentive to merge, as this will lower production costs and raise profits. This example describes a **natural oligopoly**, because demand and cost conditions make it productively efficient and most profitable for there to be just a few firms in the market.

The concept of a cost-minimizing industry structure provides a simple way of showing how scale economies in relation to the size of the market affect industry concentration. That is, when x is small and the cost minimizing number of firms is 1, then the industry is a natural monopoly. If x is very large, then the industry is naturally competitive. At intermediate values of x, we have the natural oligopoly. Thus, when scale economies increase (decrease), causing MES to shift right (left), the cost minimizing number of firms decreases (increases) and concentration rises (falls). When demand increases (decreases), the cost minimizing number of firms increases (decreases) and concentration falls (rises).

8.4.2.2 Concentration and Strategic Barriers to Entry

There has been extensive research on strategic entry deterrence, beginning with the seminal works of Bain (1956), Modigliani (1958), and Sylos-Labini (1962). To illustrate the basic idea, consider a two-stage game with an incumbent firm, a monopolist (M), and a potential entrant (PE). In the first stage, PE must decide whether to enter the market or not. In the second stage, M must decide whether to fight entry or not. Fighting means that M will expand output by lowering price if PE enters the market. This is called a predatory pricing strategy and is designed to maintain or gain a monopoly position. The question is, will M's threat to fight effectively keep PE out of the market?

The extensive form of this game is shown in Fig. 8.6. In this example, if entry does not take place, M's profits are 100. With entry and no price cutting, both firms earn profits of 30. With price cutting, both firms earn profits of 10. For price cutting to be an optimal strategy for M, it must be profitable to fight in the second stage of the game once PE enters the market. You can see from the figure that this is not true, as M's profits are 10 if it fights and 30 if it does not fight. If we assume that information is perfect and complete, PE can look forward and see that M will not fight. As a result, PE will enter because its profits are 30 if it enters (given that M will not fight) and 0 if it does not enter. Thus, the subgame perfect Nash equilibrium (SPNE) to this game is for PE to enter and M not to fight.[23]

[23] There are certain market settings where limit pricing can be effective. For example, Milgrom and Roberts (1982) show that limit pricing can effectively block entry when there is incomplete information. In their model, M has either the same or lower costs than PE, but only M knows if it is a low or a high cost producer. They show that if the probability that M is a low cost producer is sufficiently low, then it may be optimal for a high cost M to behave like a low cost M by charging a low price. This action will deter entry of PE.

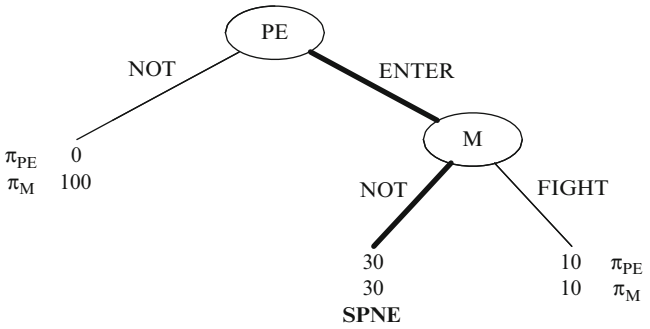

Fig. 8.6 Entry game with a potential entrant (PE) and a monopolist (M)

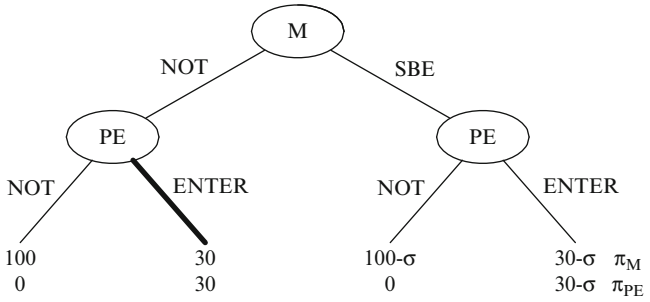

Fig. 8.7 Entry game with a strategic barrier to entry (SBE)

The reason why this strategy does not effectively deter entry is that the threat to fight is merely cheap talk and is not credible, the same problem we found in the bank robber game in Chap. 3. That is, even if M announces before play that it will fight, once PE enters it is not rational to follow through with the threat. It is not a SPNE strategy to fight. For a strategic barrier to be effective, it must be based on a threat that is credible.

One way to make such a threat credible is to formally commit to a course of action before entry takes place. M may commit to an investment that raises the sunk cost of doing business for both M and PE. Examples include investments in advertising or in research and development. That is, if M invests in research and development to improve the quality of its product, PE must do the same to remain competitive. Will an investment that raises the sunk costs of both firms deter entry?

To analyze this problem, we consider the dynamic game described in Fig. 8.7. In the first stage, M either invests in the strategic barrier to entry (SBE) or not. When M invests in the SBE, this raises the cost to both firms by $\sigma > 0$. Notice that

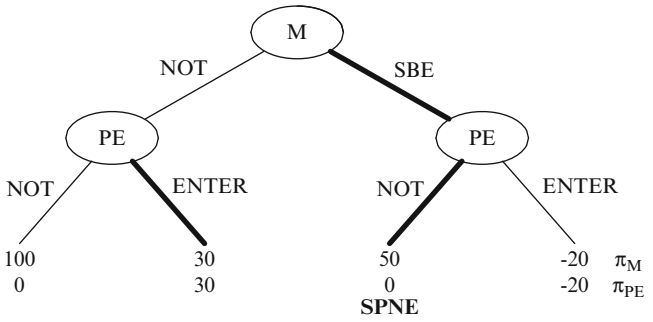

Fig. 8.8 Entry game with a strategic barrier to entry and $\sigma = 50$

M will not invest in SBE if there is no threat of entry, because M earns 100 with no SBE and earns $100 - \sigma$ with SBE. In this example, it clearly pays for PE to enter if M does not invest in the SBE. With the threat of entry, M can successfully deter entry by investing in SBE if $30 < \sigma < 70$. To demonstrate, notice that:

- The SBE fails to deter entry when $\sigma < 30$, because PE's dominant strategy is to enter when $\sigma < 30$. Thus, σ must exceed 30 to deter entry.
- It is unprofitable for M to invest in SBE when $\sigma > 70$. If M does not invest in the SBE, then PE enters and M earns 30. If M invests in SBE and $\sigma > 30$, then PE will not enter and M earns $100 - \sigma$. Thus, it will not be profitable for M to invest in SBE if $\sigma > 70$.
- This implies that M can successfully deter entry by investing in SBE if σ ranges from 30 to 70.

To provide a more specific example, consider the case where $\sigma = 50$, as described in Fig. 8.8. In this example, PE enters with no SBE and does not enter with SBE. M's payoff is 30 with no SBE and 50 with SBE. Thus, the SPNE strategy is for M to invest in SBE and PE to refrain from entry; SBE successfully deters entry and keeps concentration high.

This example shows the inefficiency that can result from a strategic barrier to entry. First, it preserves the monopolist's position, which is allocatively inefficient. Second, M invests in SBE only because it deters entry. As a result, it is socially wasteful because it is costly and serves no purpose other than to insulate the incumbent monopolist from competition.

8.4.3 Sutton's Theory of Sunk Costs and Concentration

According to Sutton (1991, 1999), sunk costs play a key role in determining industry concentration. Sutton uses the following game to illustrate the main idea. Firms compete in two stages or periods:

I. In the first stage, they must decide whether to enter the market, which requires a start-up cost that is a sunk cost.[24]
II. In the second stage, firms compete in output (or price).

There are two possible market settings. In the first, sunk costs are exogenous. This is similar to the case above where natural barriers to entry affect industry concentration. In the second, sunk costs are endogenous, an assumption that leads to considerably different results.

8.4.3.1 Exogenous Sunk Costs and Concentration

To begin, we consider a simple version of Sutton's model with exogenous sunk costs. Suppose that there is a market with n symmetric firms that produce homogeneous goods. To enter the market before competition begins (in stage I), firms must pay a set-up (quasi-fixed) cost ($\sigma > 0$) which is exogenously determined and sunk. Total revenue at the industry level (TR) is defined as n times firm i's total revenue ($p \cdot q_i$), where p is price and q_i is firm i's output. Once competition commences in stage II, a firm's price–cost margin (PCM) is defined as $(p - c)/p$, where c is the marginal cost of production, and firm i's profit is $\pi_i = (p - c)q_i$.

Sutton analyzed this model to determine the effect of sunk costs, market size, and the degree of competition on industry concentration. In the first stage of the game, firms enter the market as long as profits exceed σ. Entry continues until

$$\sigma = (p - c)q_i. \tag{8.6}$$

By multiplying and dividing through by p on the right-hand side of (8.6), we can rewrite this equation as

$$\begin{aligned} \sigma &= \frac{(p - c)}{p} pq_i \\ &= \text{PCM} \cdot pq_i. \end{aligned} \tag{8.7}$$

Recall that $\text{TR} = n \cdot pq_i$ and $1/n = \text{HHI}$ when firms are symmetric and market share is measured as a decimal. If we multiply both sides of the equality in the second line of (8.7) by n, then it can be rewritten as

$$\frac{1}{n} = \text{HHI} = \frac{\sigma}{\text{PCM} \cdot \text{TR}}. \tag{8.8}$$

[24] Sutton (2007, p. 2359) argues that if fixed costs are not sunk, then many of Sutton's conclusions are invalid because it would then be more appropriate to assume that firms play a static rather than a dynamic game. When an investment such as this is made before any output is produced, it is a quasi-fixed cost (see Chap. 2).

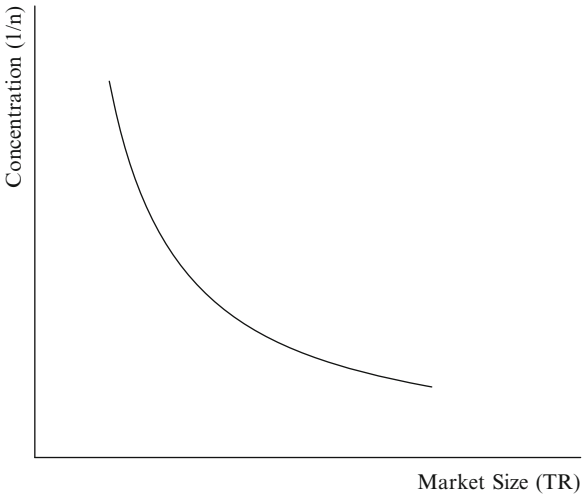

Fig. 8.9 The relationship between concentration and market size when sunk costs are exogenous

This formulation of the equilibrium has three important implications:

1. Concentration (HHI) increases with sunk costs (σ).
2. An increase in the size of the market (TR) causes concentration to fall.
3. Tougher competition, which decreases profits (PCM), causes concentration to increase.

The first two implications are consistent with those found in our discussion of natural barriers to entry. That is, as start-up costs or MES increases relative to the size of the market (i.e., market demand), concentration increases. The last implication is somewhat surprising: as firm behavior becomes more competitive, moving from monopoly (or collusive) to perfect competition, concentration rises. Sutton calls this the "toughness of competition," and his model implies that tougher competition leads to lower profits, which in turn reduces entry and raises concentration. This is a valuable contribution because it provides one mechanism by which firm behavior affects market structure.

With exogenous sunk costs and intermediate levels of competition, those between cartel and perfect competition, the level of concentration continues to fall as the market expands (see Fig. 8.9).[25] Although this relationship seems natural and appears to hold for many industries, Sutton shows that it does not always hold in markets with differentiated goods where advertising and research and development are significant features of the industry. This observation motivated his work on markets with endogenous sunk costs.

[25] Sutton assumed a Cournot model which produces an outcome that lies between cartel and perfect competition. We will discuss the Cournot model in Chap. 10.

8.4.3.2 Endogenous Sunk Costs and Concentration

Sutton's model becomes considerably more complex when sunk costs are endogenous because market structure, firm conduct, and industry performance are now determined simultaneously. The key difference in the endogenous case is that products differ in real or perceived quality, and firms can make sunk cost investments in the first period of the game to improve product quality.[26] Thus, sunk cost investments can cause firms to differ in terms of their competitive fitness. One type of sunk cost investment is research and development, which can enable a firm to gain a real quality advantage over its competitors. Alternatively, a firm may invest in advertising that improves product "image" by informing consumers of the real or perceived quality advantages of the advertised brand.

In this model, an increase in the size of the market is assumed to induce firms to boost their sunk cost investments to enhance the quality of their products.[27] We will see in later chapters that this is generally true for advertising and for research and development, as expenditures in these areas usually rise with sales.

Under these conditions, Sutton's model predicts one strikingly different result from the case of exogenous sunk costs: an increase in the size of the market will not lower industry concentration below some minimum level of concentration. We illustrate this prediction in Fig. 8.10. The intuition behind this result is as follows. As in the case with exogenous sunk costs, an increase in the size of the market raises industry profits which induces entry and puts downward pressure on industry concentration. There is an additional force at work, however, when sunk costs are endogenous. Market growth also induces firms to make investments to improve product quality, which raises sunk costs, lowers profits, and puts upward pressure on concentration. This latter effect keeps concentration from falling below a positive lower bound as the size of the market increases.[28]

---

[26] Recall from Chap. 7 that differentiation can be vertical (e.g., quality differences) or horizontal (e.g., location differences). Because assuming vertical (quality) differentiation produces such dramatically different results, we focus on vertical differentiation here (found in Sutton 1991, Chap. 3). When differentiation is horizontal, Sutton shows that the relationship between concentration and market size is less precise than for the homogeneous goods case found in Fig. 8.9 (see Sutton 1991, pp. 37–42). With horizontal differentiation, acceptable concentration and market size values include the curve and all points to the north and east of the curve in Fig. 8.9. This is called a "bounds approach," because the model provides bounds on the set of outcomes rather than pinning down a precise relationship.

[27] For example, if quality is a normal good, an increase in consumer income could increase sales and the demand for quality, which would induce firms to increase the quality of their products.

[28] In his work on research and development and sunk costs, Sutton (1999, 2007) also argues that concentration can vary, depending upon the type of technological trajectories that are characteristic of an industry. If, for example, goods are relatively homogeneous and firms compete in research and development that is designed to lower production cost (i.e., they follow a single technical trajectory), as in the aircraft industry, then concentration tends to rise over time and remain high. Alternatively, when many submarkets or niche markets exist, as in the flowmeter (i.e., devices that control the flow of gases and liquids through pipes) industry, firms may choose to compete in one or a few submarkets and pursue a proliferation of technical trajectories. This tends to keep concentration from increasing over time.

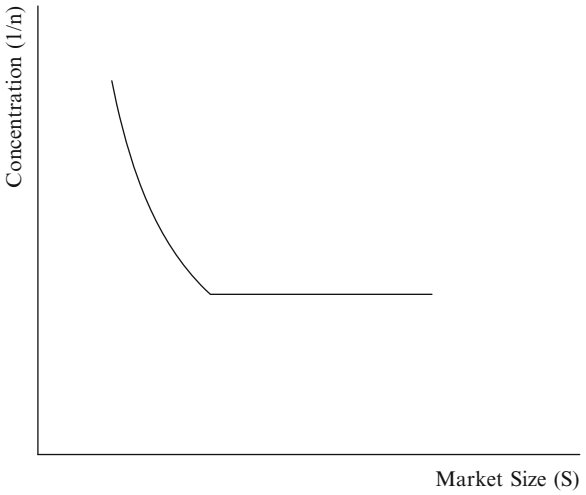

Fig. 8.10 The relationship between concentration and market size when sunk costs are endogenous

The two predictions illustrated in Figs. 8.8 and 8.9 can be summarized by **Sutton's Limit Theorem**: when sunk costs are exogenous, concentration converges to zero as market size increases; when sunk costs are endogenous, concentration converges to a lower bound that is above zero. The empirical implications of Sutton's work are clear: (1) In industries where sunk costs are exogenous, the level of concentration for a particular industry should be relatively low in large countries and relatively high in small countries. (2) In industries where there are sunk cost investments on such things as advertising and research and development, concentration should be unaffected by the size of the economy. (3) Higher levels of competition produce higher levels of concentration.

8.5 Survey and Empirical Evidence

We have seen previously that the level of industry concentration varies across industries and can change considerably over time. In this section, we summarize the empirical evidence regarding the causes of industry concentration.

Early empirical studies found general support for the hypothesis that demand and cost conditions influence industry concentration. After surveying the evidence, Schmalensee (1989) concludes:

- When concentration in a particular industry is high (low) in one country, it is frequently high (low) in other countries.
- Concentration tends to be positively correlated with MES and capital intensity.
- Outside the USA, mergers are an important cause of high industry concentration.

The reason why mergers have had a lesser effect on US concentration is that antimerger laws are generally more restrictive in the USA. The first two implications are consistent with the hypothesis that natural barriers to entry increase industry concentration. This is not surprising, since MES is likely to be similar across countries.

After reviewing the same evidence, Scherer and Ross (1990, 141) add that "actual concentration in US manufacturing industry appears to be considerably higher than the imperatives of scale economies require." This implies that strategic investments in sunk costs may also play a role, as the work of Sutton and others suggest. Kessides (1990) confirms this viewpoint, finding empirical support for the hypothesis that high sunk costs lead to high levels of industry concentration in a diverse sample of industries.

Smiley (1988) conducted a revealing survey of 293 product managers from major corporations to determine the importance of strategic entry deterrence in the USA. He found that over half indicated that entry deterring activity is as important as other strategic marketing and production decisions. In addition, firms refrained from strategic entry deterring activity when entry was unlikely, entry was inevitable, and when entry deterrence was too costly.

Smiley also tried to identify strategies that are frequently used to limit entry. The survey asked whether a particular entry deterring strategy was common practice in the industry based on a five-point scale, with five meaning frequently and one meaning never. For both new and mature industries, the survey addressed eight potentially important practices that are designed to make entry less attractive:

1. Advertising: Use advertising to create brand loyalty (brand names) and raise sunk costs.
2. Hide profits: Hide excess profits of a particular product from competitors by producing a multitude of products.[29] This applies to mature industries.
3. Brand proliferation: Offer a wide range of brands within an industry to fill all product niches. This applies to mature industries.
4. Research and development (R&D): Invest in R&D to develop new patents and increase sunk costs.
5. Reputation: Develop a reputation for competitiveness, through communication to the media or by past behavior.
6. Learning curve: Expand output today to gain experience and lower future costs. This applies to new industries.
7. Excess capacity: Build an especially large plant to meet all expected future demand.
8. Limit pricing: Choose a sufficiently low price.

The main results of Smiley's study are summarized in Table 8.9. It reports the percent of respondents who indicated that a particular strategy was frequently used in their industry (i.e., had a score of 3 or above). The figures reveal that firms in the real

[29] Stigler (1966, 227) puts it this way: "if one can conceal the profitability of his situation, entry will be slower."

Table 8.9 Frequency of strategic entry deterring strategies

	New products	Mature products
Advertising	62%	52%
Hide profits	–	59
Brand proliferation	–	57
R&D patent	56	31
Reputation	27	27
Learning curve	26	–
Excess capacity	22	21
Limit pricing[a]	9	21

[a] These are averages for static limit pricing and dynamic limit pricing.
Source: Smiley (1988).

world use a variety of methods to deter entry. The most prevalent tactics, with scores above 50%, are advertising, hiding profits, brand proliferation, and R&D. Smiley also found that R&D activity is less important in mature industries. Given their importance, much of our attention in upcoming chapters will be devoted to these strategic variables.

Regarding empirical evidence, there is considerable support for Sutton's (1991) theory of market structure. To test his theory, Sutton collected data from twenty food and beverage industries in six countries. These were divided into two groups: those with homogeneous goods and little or no advertising and those with moderate to high levels of advertising.[30] As Table 8.8 indicates, concentration is generally higher for the advertising-intensive group.

Sutton also used regression analysis to determine the effect of market size on concentration for these two groups. The simplest version of the model is presented below:

$$\mathrm{CR}_4 = \beta_0 + \beta_1 \ln\left(\frac{\mathrm{TR}}{\sigma}\right) + \beta_2 x, \tag{8.9}$$

where the βs are regression parameters, TR is industry sales or total revenue, σ measures start-up costs (i.e., the size of an efficient plant), and x is a vector of other control variables.[31] Sutton's theory predicts that β_1 will be negative for markets with homogeneous goods and 0 for advertising-intensive markets, which is exactly what he found. The regression estimate of β_1 was -0.187 ($t = 3.2$) for homogeneous-goods markets and was –0.02 ($t = 0.63$) for advertising-intensive markets. Thus, endogenous sunk costs associated with advertising appear to substantially diminish the effect of the size of the market on industry concentration.

[30] Sutton (1999) also finds support for his theory when research and development expenditures are the primary source of sunk costs.

[31] Control variables include dummy variables for countries and industries; β_2 is a vector of parameters conformable to x. For further discussion, see Sutton (1991, Chaps. 4 and 5).

More recently, Ellickson (2007) analyzed Sutton's model using data from regional US supermarkets. Rather than competing in advertising, Ellickson found that supermarkets competed by offering a greater selection of products. If Sutton's model is correct, an increase in the size of the market should induce firms to build larger stores and offer greater product variety. This in turn would raise sunk costs and keep concentration from falling as the market expands. Ellickson discovered that as individual markets grew, concentration (measured by CR_1, CR_2, CR_4, CR_8, CR_{20}, and HHI) remained virtually unchanged.

Symeonidis (2000, 2001) tested the implication of Sutton's model that greater competition leads to higher industry concentration. He analyzed a natural experiment in the UK in the 1960s when the laws against cartel behavior were strengthened. By reviewing data from a general class of manufacturing industries, he found strong support for Sutton's work. Stiffer laws against cartels resulted in greater price competition, which increased industry concentration by diminishing the number of firms through exit or merger.[32]

Although there is general support for Sutton's work, we should keep one caveat in mind. As Sutton (1991, Chap. 9) points out, sunk costs are not all that matter in the evolution of market structure. History and the idiosyncratic characteristics of an industry may also have influence. One example is when a firm has a first-mover advantage and gains a dominant position, resulting in high levels of concentration. For example, Alcoa gained an early advantage by being the first to acquire aluminum ore deposits in the USA. Similarly, Anheuser-Busch benefitted from locating its first brewery on land with deep caves that could be used to store beer at cool temperatures. This gave the company a strategic advantage before the advent of refrigeration. In any case, the evidence clearly shows that high sunk costs can be an important contributor to high industry concentration.

8.6 Summary

1. Market structure refers to the way in which a market is organized. Markets fall into one of four broad categories: perfect competition, monopoly, monopolistic competition, and oligopoly.
2. An **oligopoly** market consists of just a few competitors in which products are homogeneous or differentiated. The key feature of this market structure is strategic interaction, in that a firm's profits depend on the actions of rival firms as well as its own actions. Thus, game theory is used in oligopoly modeling. Oligopoly is the most common market structure in the US economy.

[32] Other studies include Robinson and Chiang (1996) for a sample of US consumer goods industries, Matraves (1999) for the global pharmaceutical industry, Lyons et al. (2001) for a sample of industries in the European Union, and V. Tremblay and C. Tremblay (2005) for the US brewing industry. See Sutton (2007) for a more extensive survey of the empirical evidence.

3. A key element of market structure is industry concentration, which is described by the number and size distribution of firms. Competitive and monopolistically competitive industries have many firms of equal size, and a monopoly has just one firm. There are only a few firms in an oligopoly market, and firms may or may not be symmetric in size.
4. A **concentration curve** provides a visualization of industry concentration. It plots the market share of the largest firm, the two largest firms, the three largest firms, and so on for all firms in the industry. A linear concentration curve implies that firms are of equal size. Fewer firms and a more unequal distribution of firm size shift the curve up and to the left, implying a greater level of industry concentration.
5. It is useful, especially in empirical work, to identify concentration with a single index. Ideally, such an index should increase with the number of firms and with the extent of inequality in the distribution of market shares. We have discussed three indices of industry concentration:

    1. The **number of firms** (n). This is an unsatisfactory index unless firms are symmetric.
    2. The ***k*-firm concentration ratio** (CR_k), which measures the market share of the largest k firms in the industry. The main advantage of this index is that it is easy to calculate. However, it provides no information about the distribution of market shares among the largest k firms, and it ignores firms outside the largest k firms. Thus, it does not always provide a ranking of industry concentration that is consistent with a concentration curve.
    3. The **Herfindahl–Hirschman index** (HHI), which equals the sum of squared market shares of all firms in the industry. Although the calculation of HHI requires a great deal of data, it has the desirable qualities of increasing with the number of competitors and with the inequality of the distribution of firm sales. When market share is measured as a decimal, $HHI = n\sigma^2 + 1/n$, where σ^2 is the variance of firm market share. This implies that when firms are symmetric (i.e., $\sigma^2 = 0$), $HHI = 1/n$. The relationship that $n = 1/HHI$ is called a **numbers equivalent** because it implies that a given value of HHI can be translated into a number of equal sized firms.

6. A **dominant firm** has a larger market share than its competitors and typically takes a leadership role in choosing price or output. Dominance can result from producing a superior product or from producing at lower cost than competitors.
7. Experts use CR_4 and HHI measures of concentration to distinguish between competitive and oligopoly markets. In terms of CR_4, an industry is classified as oligopolistic once CR_4 reaches 40%. The Department of Justice and the Federal Trade Commission use the following delineation:

    - An industry is classified as unconcentrated when HHI is less than 1,000.
    - An industry is classified as moderately concentrated when HHI ranges from 1,000 to less than 1,800.
    - An industry is classified as highly concentrated when HHI is greater than or equal to 1,800.

8. A critical step in measuring concentration is to properly define the market. The relevant product market includes all products that are close substitutes in consumption and production. Identifying a market also requires the proper definition of the geographic boundary, as markets may be local, regional, national, or international.
9. **Aggregate concentration** is defined as the market share of total US sales that are produced by the largest corporations. Although there are concerns that massive corporate size may provide firms with political and economic power, aggregate concentration has remained relatively stable over the last 50 years.
10. A review of concentration in US industries reveals the following:
    - Concentration varies considerably from industry to industry.
    - Although concentration is relatively stable over time in some industries, it has changed dramatically in others.
    - Across countries, when industry concentration is high (low) in one nation, it tends to be high (low) in other nations.
11. A number of forces cause industry concentration to be high. One is described by **Gibrat's Law**, which says that luck or random shocks to firm growth rates can cause the distribution of firm size to become more skewed, thus raising industry concentration. Traditionally, **barriers to entry** are viewed as the fundamental cause of high concentration. A barrier to entry is defined as any limitation on entry that keeps the long-run equilibrium number of firms below the competitive number. Barriers to entry are classified into three groups: natural barriers, legal barriers, and strategic barriers. **Natural barriers** exist when demand and cost conditions limit the number of firms. **Legal barriers** include government regulations that legally restrict entry. In general, natural and legal barriers are exogenously determined. **Strategic barriers** include all predatory actions of firms that limit entry. These are clearly endogenous barriers to entry.
12. The **cost-minimizing industry structure** is defined as the number of firms needed to produce industry output at minimum cost. When industry cost minimization occurs, the industry is productively efficient. In the case of oligopoly (monopoly), the cost-minimizing industry structure is normally determined by natural barriers to entry. A **natural oligopoly** (monopoly) occurs when the cost-minimizing industry structure is just a few firms (one firm).
13. High sunk costs can be a barrier to entry because a sunk cost represents an expenditure that cannot be recovered when the firm exits the industry.
14. Sutton (1991) developed a model where sunk costs have a critical effect on industry concentration. His model predicts that concentration increases with sunk costs and the vigor of competition. Sutton's model also shows that the effect of sunk costs on industry concentration will be different, depending on whether the sunk costs are exogenous or endogenous. These results are summarized in the **Sutton Limit Theorem**: when sunk costs are exogenous, concentration converges to zero as market size increases; when sunk costs are endogenous, concentration remains above a lower bound when market size increases. Although

sunk costs are important, Sutton also points out that history and the idiosyncratic characteristics of an industry may also influence industry concentration.

15. Survey evidence indicates that the most effective entry deterring strategies are advertising, hiding profits, brand proliferation, and R&D. As one might expect, R&D activity is less prevalent in mature industries.
16. The empirical evidence regarding the main causes of high industry concentration is generally consistent with economic theory. The main conclusions are:

    - In a particular industry, when concentration is high in one country, it is frequently high in other countries, especially in industries with exogenous sunk costs.
    - Concentration tends to be higher in markets with high natural barriers to entry, such as when MES and capital costs are high relative to the size of the market.
    - Mergers are an important source of concentration, especially outside the USA.
    - Concentration tends to be high in markets with high sunk costs and when firms invest in strategic barriers to entry, which generates endogenous sunk costs.

8.7 Review Questions

1. Define industry concentration. Explain how a concentration curve can be used to describe industry concentration. Can a concentration curve be (strictly) convex from below?[33] Explain.
2. Regarding an index of industry concentration.

    A. Describe the characteristics of an ideal index of industry concentration.
    B. Do the three indices of industry concentration described in the book (n, CR_4, and HHI) meet these ideal characteristics?
    C. How is HHI related to n and to CR_4?

3. Explain what is meant by an economic market. How do product and geographic boundaries play a role in your definition? Why is it important to use the correct economic market when constructing an index of industry concentration?
4. Describe what is meant by aggregate concentration. How has aggregate concentration in the USA changed in the last half century? Interpret the mean value of CR_{50} in Table 8.4. Why might high aggregate concentration be a social concern?
5. Table 8.6 shows that HHI is 350 for frozen fruits, juices, and vegetables, 467 for cement, and 2,449 for breakfast cereal. Why is this measure of concentration in the cement market inaccurate? Why do you think that concentration is low for the frozen food industry and high for the breakfast cereal industry?
6. Assume a market where firms produce homogeneous goods and are symmetric (i.e., each firm produces the same amount of output in equilibrium). The long-run

[33] A curve is convex when it lies above any tangent line to the curve.

average cost (AC) curve is U-shaped, and total demand is 120 (million units) when price equals minimum long-run average cost.

A. If minimum efficient scale (MES) is 10, what is the cost-minimizing number of firms (n^*)?
B. If MES is 11, what is the cost-minimizing number of firms? How will your answer change if AC is relatively flat to the right of MES.

7. Explain how strategic barriers to entry are different from natural barriers to entry. Provide one example of each.
8. Use Sutton's model to explain how concentration is determined when sunk costs are exogenous and when they are endogenous. Use an increase in industry sales or revenues (TR) to explain your answer.
9. Sutton (1991, Chap. 2) developed another model with exogenous sunk costs (σ) where equilibrium profits for firm i are $\pi_i = \text{TR}/n^2 - \sigma$.

A. Explain how TR, n, and σ affect firm profits.
B. What will be the equilibrium number of firms in this market?

10. Provide a brief summary of the empirical evidence regarding the main causes of high industry concentration.