Dennis W. Carlton, *Barriers to Entry*, *in* 1 ISSUES IN COMPETITION LAW AND POLICY 601
(ABA Section of Antitrust Law 2008)

Chapter 25

BARRIERS TO ENTRY

Dennis W. Carlton[*]

This chapter analyzes the concept of barriers to entry. It explains that the concept is a static one and explores the inadequacy of the concept in a world with sunk costs, adjustment costs, and uncertainty. The static concept addresses the question of whether profits are excessive. The more interesting and relevant questions are how fast entry or exit will erode profits or losses, and how do the bounds that entry and exit place on price vary with uncertainty and sunk cost. Intuition based on the static concept of barrier to entry can be misleading in many industries.

1. Introduction

The concept of barriers to entry has been a barrier to economists' understanding of industrial structure and has misled courts and regulatory agencies repeatedly as they attempt to use the concept in antitrust cases or regulatory proceedings. There are two primary reasons why it has proved so confusing a concept. First, the theoretical underpinnings for the concept arise from the structure-conduct-performance literature, which itself has been shown to suffer severe theoretical problems. Second, a large part of the confusion has arisen because commentators are often unclear about the precise consequences of a barrier to entry. For example, is price too high, profit too high, entry too slow, social welfare too low, or all of these? If the point of defining barriers to entry is to identify some (exogenous) conditions that imply social harm, one should not define barriers as conditions that cause social harm, unless one can identify the conditions ex ante before solving for the equilibrium. Otherwise, such a definition serves little purpose.

The works of George Stigler[1] and Harold Demsetz[2] in the late 1960s clarified much of the confusion surrounding the meaning of "entry barriers." However, those works, like most work in industrial organization, ignored dynamics and, in particular, uncertainty and adjustment costs. There is often a confusion of adjustment costs with a barrier to entry, even when competition prevails. More generally, industrial organization economists have tended to ignore adjustment costs and therefore think in terms of short run and long run only—useful pedagogical tools, but ones that are often

[*] Graduate School of Business, University of Chicago; National Bureau of Economic Research; and U.S. Department of Justice. This chapter does not necessarily reflect the views of the Department of Justice. The author thanks Elizabeth Bailey, Dale Collins, Robert Gertner, Gregory Pelnar, Sam Peltzman, Robert Pindyck, Richard Posner, Thomas Ross, and Hal Sider for helpful comments. This chapter draws upon and extends Dennis W. Carlton, *Why Barriers to Entry Are Barriers to Understanding*, 94 AM. ECON. REV. 466 (2004). The author also would like to thank the *American Economic Review* for allowing him to incorporate that paper into this chapter.

[1] GEORGE STIGLER, THE ORGANIZATION OF INDUSTRY (1968).

[2] Harold Demsetz, *Why Regulate Utilities?*, 11 J.L. & ECON. 55 (1968).

inadequate to address practical antitrust and regulatory problems. Furthermore, industrial organization economists have generally ignored the consequences of uncertainty. By introducing uncertainty into a dynamic model, new insights emerge about how entry and exit constrain price in an industry. Many of these insights are contrary to the intuition one obtains from models with no uncertainty. For example, one may observe no entry in a competitive industry earning significant short-run profits over extended periods.

Although there is undoubtedly disagreement among economists as to what constitutes a barrier to entry, that disagreement does not always lead to great misunderstanding among economists as to the predictions of industry outcomes. Economists are much better at figuring out the market equilibrium than agreeing on and applying definitions. But definitions and their application can have enormous effects if they are used as tools of analysis in antitrust cases and in regulatory proceedings. Then, disagreement over whether a barrier exists or not can matter in terms of the disposition of an antitrust trial or regulatory proceeding, even if there may be no disagreement among economists about the effect of the disputed barrier on a market's equilibrium.

This chapter is organized as follows. Section 2 explains the deficiency in relying on the structure-conduct-performance model to understand entry barriers. Section 3 describes how Stigler and Demsetz clarified the notion of entry barriers in a static world. Section 4 explains why dynamic models are needed to understand the entry process and to understand what can slow down or eliminate a firm's incentive to enter. Section 5 discusses the consequences of introducing uncertainty into a dynamic analysis involving sunk costs. Section 6 discusses briefly the empirical evidence on entry and exit. Section 7 relates the analysis of entry and exit to antitrust and regulatory concerns. Section 8 presents a conclusion.

2. Bain's barriers to entry

Joe Bain deserves credit for trying to find presumably exogenous factors of industry structure that influence how competition occurs and that prevent price from reaching the competitive level.[3] Bain's investigations led him to identify several factors as barriers to entry, such as scale economies, large capital requirements, product differentiation, and cost advantage. These barriers protect a firm from entry and thereby enable it to enjoy above normal rates of return.

The problem with Bain's analysis of entry barriers is not his definition (namely, entry conditions allowing for an elevated long-run price), but his failure to articulate a consistent theory whereby the factors he identifies as entry barriers such as scale economies, large capital requirements, and product differentiation lead to such an elevated price. Bain's analysis makes most sense if one has a view of the world in which barriers determine the number of firms which, in turn, determines the competitiveness of the industry and thereby determines each firm's rate of return. This structure-conduct-performance view of the world is, alas, too simple. The number of firms is typically determined by a decision to enter based on profitability. Profitability

[3] JOE S. BAIN, BARRIERS TO NEW COMPETITION: THEIR CHARACTER AND CONSEQUENCES IN MANUFACTURING INDUSTRIES 3 (1956).

for any given number of firms, however, is determined not just by exogenous factors such as costs, but also by price, which will be determined by the "vigor of competition" or, in game theory terms, by the competitive game being played. The effect of a particular barrier to entry on price determination will depend on the nature of this competitive game. There is no reason to assume that this game is the same across different industries. Economists have made little progress in explaining how this competitive game depends on exogenous factors. It is possible to show how industries with the exact same structure (e.g., cost conditions, demand conditions) will have very different outcomes depending on the particular form of competition in the industry. For example, John Sutton's pathbreaking work[4] convincingly demonstrates that industries where competition is very vigorous will be more highly concentrated than those where competition is not as vigorous.[5] High concentration, far from being an indicator of a lack of competition, can indicate precisely the reverse. In those industries, there may be large firms, but the large size will not be associated with high price. Therefore, the effect of a structural industry factor on the equilibrium price will vary enormously across industries, and what is a "barrier" in one industry may have no effect on price in another.

Sutton's point can be easily seen by considering the following. For simplicity, assume that each firm has the identical cost technology, which consists of a fixed cost plus a constant marginal cost. Imagine an industry that in one country (country A) is described by quantity (i.e., Cournot) competition and free entry but in another country (country B) is cartelized with free entry into the cartel (say, by regulation). The price in country B and total industry variable profits will exceed that in country A, yet with a fixed cost of entry, country B can support more firms than country A. Hence with free entry, industry concentration in country A will exceed that in country B. Price is higher in the country with lower concentration.

Unless one would include the vigor of competition in the list of factors defining entry barriers, it is misleading to treat the number of firms as determined by entry barriers, and it seems an odd use of language to term vigor of competition as an entry barrier. Indeed, the example illustrates the difficulty with treating the number of firms as determined by entry barriers alone.

As explained earlier, the number of firms, their size, and the equilibrium price is determined by more than just the factors Bain claimed. But one can go further, as Sutton has, and claim that several of the factors that Bain treats as determinants of the industry equilibrium such as product differentiation and cost advantage are themselves not exogenous because they can be altered by investment. Hence, they are, in equilibrium, determined by the same economic forces that determine industry concentration. For example, firms can compete against each other by investing in the development of new products, in the promotion of the product, or in the reduction of costs. All these features are determined in equilibrium together with industry

[4] JOHN SUTTON, SUNK COSTS AND MARKET STRUCTURE (1991) [hereinafter SUTTON, SUNK COSTS]; JOHN SUTTON, TECHNOLOGY AND MARKET STRUCTURE (1998) [hereinafter SUTTON, TECHNOLOGY].

[5] The "vigor of competition" can be precisely defined. Industry A is more vigorously competitive than industry B if, all else equal, price is lower for industry A for any given number of firms. *See* SUTTON, SUNK COSTS, *supra* note 4, at 33, where he discusses "toughness."

concentration. One can show in these models that as markets grow in size, the industry structure that can emerge is not one of atomistic competition with constant quality but rather one where concentration remains high but product quality increases. Therefore, competition along nonprice dimensions can explain why concentration does not necessarily diminish as industries grow. The significance of this point cannot be overstated. Models that focus on only price competition may fail miserably to correctly predict industry concentration and consumer welfare when there are other product dimensions along which competition occurs. This is likely to be particularly true in industries requiring investment and creation of new products. It is no coincidence that many of the most controversial antitrust and regulatory cases have arisen in high-technology industries (e.g., computers and telecommunications) where competition in research and development and new products is paramount.

Thus, although Bain deserves credit for identifying what are interesting facts about an industry, these facts are not necessarily exogenous and do not alone determine the number of firms, which in turn does not alone determine price. Bain's quest to identify barriers lacks theoretical rigor.

3. Stigler and Demsetz

The confusion surrounding the meaning of "barriers to entry" often results because the precise consequence of having an entry barrier is unclear. If there are such "barriers," are rates of return too high? Is the existence of such barriers socially undesirable, and therefore, is it proper to use antitrust laws (or legislation) to attack the problem? Although Bain was trying primarily to answer the first question, many have focused on the second. The two questions are very different, and the first question is not the right one for antitrust or regulation to focus on. Moreover, the concept of (static) entry barriers may not be particularly appropriate for answering the second question.

Stigler clarified matters by defining an entry barrier as "a cost of producing (at some or every rate of output) which must be borne by firms which seek to enter an industry but is not borne by firms already in the industry," that is, a cost advantage that an incumbent firm enjoys compared to entrants.[6] With such an advantage, the incumbent firm can permanently elevate its price above its costs and thereby earn a supracompetitive return.[7] This means that, if the incumbent firm has to spend $1 million annually to maintain the reputation of its brand and thereby the loyalty of its customers, then as long as a new entrant could do the same for the identical physical product, there is no reason to expect the incumbent firm to earn supracompetitive returns. Stigler paid no attention to dynamics or sunk costs in his discussion and focused implicitly only on the long-run steady state. A sunk cost, like a fixed cost, does not vary with output, but unlike fixed costs, is not recoverable if the firm shuts down. For example, annual rent is

[6] *See* STIGLER, *supra* note 1, at 67. The definition of a long-run barrier to entry advanced by Carlton and Perloff is Stigler's definition but limited to the long run. *See* DENNIS CARLTON & JEFFREY PERLOFF, MODERN INDUSTRIAL ORGANIZATION 77 (2005).

[7] Note that a scarce input (e.g., managerial talent) will earn a rent for the owner of that scarce resource but does not necessarily create an inefficiency as long as each firm operates where price equals marginal cost.

a sunk cost if the firm cannot re-lease its office space pursuant to its lease if the firm ceases operations during the year but is not a completely sunk cost if the firm can recover some or all of the annual rent by subletting its office space to another firm. Notice how, for any annual rent, the cost of failure is larger when the annual rent is sunk rather than fixed but not sunk.

Demsetz's classic article further clarified matters by considering a model in which it is efficient to have only one firm, an extreme example of scale economies.[8] As long as an entrant and the incumbent are on equal footing to bid for customers, there is no reason to expect the winner to earn supracompetitive returns. Demsetz's analysis, like Stigler's, pinpoints symmetry as the key to answering the question: what determines whether a firm can earn excess returns? Demsetz's work provided a foundation for contestability,[9] in which costless entry and exit places all firms in a symmetric position.

Although Stigler's definition of barrier as a differential cost is concise and unambiguous, it does raise the question of why it should be called a barrier. Why not call it "differential cost advantage?" This may seem overly pedantic, but introduction of unnatural use of language can lead to confusion. Consider, for example, an industry where the government restricts the number of firms to 100. The government issues 100 licenses to operate and sells them in an open market. The entry restriction is likely to be inefficient, but as long as all firms have access to the (artificially) scarce license at the market-clearing price, there is no entry barrier according to Stigler's definition. All firms earn a normal rate of return. Yet there is a restriction to entry. It seems to mangle the English language to refuse to call this entry restriction a barrier to entry. Using language in an unnatural way invites confusion in antitrust and regulatory proceedings.

4. Dynamics

The usual discussions of barriers to entry typically focus on the long run and ignore adjustment costs. In the short run, the concept of an entry barrier is not meaningful (since, by assumption, entry is not possible). But the long run is only of interest because economists often slip into ignoring dynamics and go back to the simple models of short and long run. But as a practical matter, the long run may be of little if any interest. It may take so long to get there that the persistence of supracompetitive profits in the transition period as the market adjusts to a long-run equilibrium turns out to be the fact of practical importance, not that these excess profits going forward will be eliminated in some far-off future year.

Now introduce the notion of adjustment costs, i.e., costs incurred as the firm alters its output. Adjustment costs, once incurred, are sunk costs. Those adjustment costs, together with industry characteristics (including the competitive game), will influence the speed with which equilibrium adjusts over time. It is not typically a helpful thought experiment for public policymakers to ask what would occur if adjustment costs were zero. That is a bit like asking if wages were zero, what would the new equilibrium be? Because adjustment costs are typically positive and because they are not a market

[8] Demsetz, *supra* note 2.

[9] *See, e.g.*, WILLIAM BAUMOL, JOHN PANZAR & ROBERT WILLIG, CONTESTABLE MARKETS AND THE THEORY OF INDUSTRIAL STRUCTURE (1982).

imperfection (any more than a wage is a market imperfection), one is likely to obtain misleading insights into policy by ignoring adjustment costs. There is no reason to call adjustment costs an entry barrier.[10] Trying to use "barriers to entry" to refer to both the factors that influence the time it takes to reach a new equilibrium and to whether there are excess long-run profits is confusing.

Once dynamics enter the picture, there can be all sorts of strategic behavior that can advantage one firm over another. The source of any successful strategic behavior must ultimately be traceable to an asymmetry among firms. What game theory makes clear is that in the presence of sunk costs credible commitments can be made, and these commitments can influence the equilibrium. An asymmetry can arise because one firm is in a market before another firm and can therefore act to make binding commitments before others. For example, building a specialized plant with a large capacity in advance of others may be a way to make a credible commitment to produce large outputs, and this investment may advantage the firm making the investment by deterring rivals from investing. In such situations where strategic behavior can occur, price may remain higher than one would expect based on models with no strategic behavior.

By focusing on dynamics, one can now ask different, and more relevant and detailed, questions than those suggested by prior thinking dominated by concepts of the short and long run. One can ask not only whether price will eventually converge to the competitive level, but also how long it will take before price reaches the level to which it eventually converges. In response to, say, a merger that winds up raising price by 10 percent, how much of that price increase will be eroded by entry in two years or five years? If an incumbent firm uses exclusive five-year contracts with its distributors and 20 percent of them expire each year, how long will take for price to adjust in response to a surge in demand, compared to the case of three-year contracts? These are much more detailed questions to answer and are of more practical importance than the one implicitly posed by Bain or Stigler in defining entry barriers (i.e., what conditions allow one firm to earn excess long-run profits).

5. Uncertainty, sunk costs, and dynamics

The inadequacy of the (static) concept of barriers to entry becomes even more striking when uncertainty is introduced into a dynamic setting in the presence of sunk costs. Indeed, the introduction of uncertainty can alter some of one's basic intuitions about entry and exit even in competitive markets. It is a topic that unfortunately has not received much attention in the world of antitrust.[11]

5.1. Certainty and no adjustment costs in the long run

Consider the standard theory under certainty. Imagine that there are two periods, period 1 (the short run) and period 2 (the long run). All firms are assumed identical, no

[10] One could refer to adjustment costs as an "impediment to entry," but one must then be careful not to confuse "impediment" and "barrier."

[11] It is also a complicated topic. For a more detailed treatment, see, e.g., AVINASH K. DIXIT & ROBERT S. PINDYCK, INVESTMENT UNDER UNCERTAINTY (1994); Robert S. Pindyck, *Sunk Costs and Real Options in Antitrust Analysis*, which appears as Chapter 26 in this book.

entry can occur in the short run, though exit can occur. In the short run, the number of firms is fixed at some number N_1. If firms have the standard U-shaped average cost (AC) and average variable cost (AVC) curves with minimums respectively at C_2 and C_1 (see Figure 1), then the following equilibria occur. In the short run, price, *P*, is determined by the intersection of supply (which is the horizontal sum across N_1 firms of their individual supply curves) and demand. Price can never be below C_1, because a firm already in the industry would rather produce nothing than make a loss of $C_1 - P$ on each unit.[12] If the price is below C_2, firms are losing money (once fixed costs are considered). As a result, firms will not willingly want to reinvest to remain in the industry and will eventually exit when the opportunity arises. Enough firms will exit until the price is driven up to C_2. Since all the firms are identical by assumption, the identity of the firms that exit and those that remain is not determined. In the long run, price will equal C_2, at which point there is an incentive for neither entry nor exit.

This description of equilibrium, though standard, is inadequate because it fails to explain why in the short run price does not equal C_2. More specifically, the description fails to explain why N_1 firms are present in the short run. One explanation is that there were prior periods in which price did not equal C_2 and that one is observing an adjustment over time to altered but foreseen demand changes. That explanation was already discussed in the previous sections involving adjustment costs.

A second explanation is that price is not perfectly predictable, i.e., firms were surprised, and price in period 1 was, say, lower than expected. But this observation highlights the need to model explicitly the effect of uncertainty on entry or exit decisions. The next subsections focus on this second explanation, explore how uncertainty and sunk costs influence the decision of whether to enter or exit, and show that these two factors have profound implications on the timing of entry and exit. The effects here are in addition to the effect of adjustment costs on timing.

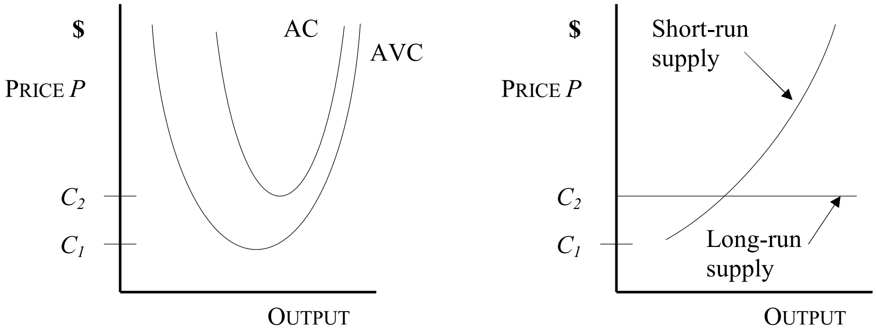

Figure 1.
Standard theory under certainty.

[12] C_1 is sometimes called the shutdown point. For simplicity, assume that in the short run all fixed costs are sunk in the sense that they are not recoverable if exit occurs.

Return to the simple model consisting of two periods: period 1 (the short run) and period 2 (long run). Suppose that C_1 = \$4 and C_2 = \$5. The standard model predicts that price can never be below \$4 and can be above \$5 only in the short run, since a price above \$5 will induce entry in the long run. Previous sections have already discussed how the short run is really a simplification and that to be more accurate one should talk in terms of speed of adjustment. Even with that caveat, the insight of the model is not to expect price to stay above \$5 without inducing entry.

5.2. Uncertainty and sunk cost with no time dependence

This standard model, though, fails to account for the consequences of the unexpected shift in demand that can cause price to be below C_2 in period 1. Presumably these shifts in demand happen all the time and if so price will vary and decisions about entry in period 1 will be based on the expectation of future prices. So, for example, the decision of a firm to become a car manufacturer will depend on the firm's expectation of future prices for cars. Similarly, the decision of a farmer whether to plant corn or wheat will depend on his expectation of the future prices of corn compared to wheat.

Let us explicitly model this uncertainty about future prices and see how it influences the entry decision. Suppose firms must decide to invest and enter at the beginning of each period before observing actual price, which is determined by the intersection of the supply curve with the uncertain demand curve. After entering and observing price, firms decide how much to produce. Suppose further that the investment is sunk and must be made at the beginning of each period (i.e., investments do not last more than one period). It is not possible to hold inventory from one period to the next. Notice that there is now no difference between period 1 and period 2 because there is no effect of one period upon another (i.e., no time dependence). The economic characteristics of each period are the same. It follows that a firm will invest and enter provided that the firm's expected profits equal zero. This means that entry will not occur in period 2 unless expected profits in period 2 are zero or above, even though price in period 1 exceeds \$5. Moreover, each firm produces more when price is high than when price is low. This means that the average price received (the price weighted by relative quantity produced) exceeds the expected price (which is the unweighted price). It is also true that one can show that this average price exceeds C_2 (\$5), while the expected price is below C_2 .[13]

[13] *See* Eytan Sheshinski & Jacques H. Dreze, *Demand Fluctuations, Capacity Utilization, and Costs*, 66 AM. ECON. REV. 731 (1976). With free entry, expected profits equal zero; hence total expected revenues equal total expected cost. This implies that average price received equals average cost incurred. But with a standard U-shaped average cost curve with a minimum average cost of C_2, the average cost incurred will exceed C_2 unless output is constant at the cost-minimizing level. But with fluctuating demand, it is not profit maximizing for output to be constant. Hence, average cost incurred will exceed C_2. When average price received equals average cost incurred, profits are zero, so it follows that average price received exceeds C_2. Consider a firm that produced that output whose average cost is C_2. That firm's profit would equal expected price minus C_2 on average. But that must be negative since it is profit maximizing to vary output as price varies over time, and in equilibrium the profit-maximizing firm earns zero profits with free entry. *See id.*

In summary, when the investment required for entry each period must occur before price can be observed, and when future demand is uncertain causing price to fluctuate, entry occurs only if expected profits are nonnegative. The observed price will fluctuate from a floor of $4 to higher prices. The average price paid will exceed $5, while the expected price will be below $5. These results arise in a competitive model with free entry at the beginning of each period. Entry does not eliminate the price fluctuations. It simply determines the equilibrium number of firms needed to drive expected profits to zero. If the demand uncertainty in period 1 is independent of the demand uncertainty in period 2, the equilibrium number of firms in period 1 and 2 will be the same. There will be no additional entry next period even if price exceeds $5 this period. Period 1 and period 2 are not connected in the sense that what happens to price in period 1 has no effect on what happens to price in period 2.

5.3. Uncertainty and sunk cost with time dependence

In many markets, there is a relationship over time in the equilibrium in periods 1 and 2 so that the price in period 1 is useful for predicting the price in period 2, in contrast to the model just presented. For example, because it takes time to expand a manufacturing facility, the capacity of a plant to produce in period 2 will depend on how big the plant is in period 1. As already discussed, adjustment costs can cause such a dependency if investments in period 1 can influence the production capacity in period 2. That element is missing in the model under uncertainty just presented. Another important reason for a time dependency in the equilibrium is a time dependency in the underlying uncertainty. For example, firms may be using a new technology whose costs are uncertain initially but eventually become known. In other markets, price may initially be highly uncertain but that uncertainty may change over time, or the expected price could change over time. In these markets, the decision to enter or exit is more complicated than in either the simple deterministic model or even the previous model with uncertainty. The reason is that now the decision is not just whether to invest to enter, but also when to do so. By delaying a decision, the firm preserves the option of making the decision in the future when there may be more information.[14] The combination of long-lasting investment plus time dependency of uncertainty can lead to different patterns of prices and entry or exit than one might otherwise expect.

To illustrate these ideas, consider a modification of the model of the previous subsection. The previous subsection used a model in which firms had to invest, then observed price. After the price was observed, the firm produced, sold its output, and its investment was used up. The situation would then repeat itself in period 2. There was no gain to waiting to enter because there was no gathering of new information in period 2 and investments made in period 1 had no effect in period 2. Now, change two features of the model. First, assume that the investment lasts two periods, not one. This induces a time dependency in the model, so that equilibrium in period 2 depends on what happens in period 1. Second, suppose that by the beginning of period 2, firms have a better idea of what price might be or by how much price is likely to fluctuate compared

[14] Notice that in the model of the previous section, there was no cost to delay. Because there was no time dependence, the decision to invest in period 2 was unaffected by decisions in period 1.

to their knowledge at the beginning of period 1. Firms can enter either at the beginning of period 1 or period 2. In equilibrium, the expected profit of a firm that enters in period 1 is zero and so too for a firm that enters in period 2. The advantage of entering in period 1 is that the firm can earn revenues for two periods. The disadvantage is that the period 1 price may be low, in which case the firm will do poorly in both period 1 and 2 because a low price in period 1 indicates that price will also be low in period 2. Had the firm waited and observed a low price in period 1, it would have decided not to enter in period 2.

Suppose the demand curve in period 1 is either $100 - P$ or $90 - P$ with equal probability. The demand in period 2 is whatever it is in period 1. For an investment of $10, a firm can enter in period 1, and produce (at zero cost) one unit in period 1 and one unit in period 2. If a firm chooses to enter in period 2, it must pay $7, though it will produce output of one unit for only one period.

In period 1, 89 firms enter.[15] With probability ½, the price in period 1 is $11 and this high price induces four additional firms to enter in period 2 to drive price to $7 in period 2. With probability ½, the price in period 1 is $1, and this low price induces no additional firms to enter in period 2, so the price in period 2 remains at $1. Notice that a firm that enters in period 2 earns zero profits. A firm that enters in period 1 also earns zero profits since its expected revenue over the two periods is just ½ (11 + 7) + ½ (1 + 1) or $10, which exactly equals the initial investment cost of entering.

The equilibrium has the following characteristics:

1. The decision to enter depends on the current state of information and that information changes over time. At the beginning of period 1, price is uncertain. At the beginning of period 2, price is known.
2. The timing of entry will involve trading off the benefit of early entry (earning expected profits in period 1) versus the benefit of delay (avoiding a bad price outcome and being stuck with it for a while). An early entrant foregoes the option to wait and enter in period 2.
3. In a competitive industry with free entry, the option value to waiting equals zero (i.e., expected profits equal zero regardless of whether the firm enters in period 1 or 2). If there are only a limited number of firms that can enter, there will be positive profits but the expected value of waiting will equal zero (i.e., firms earn the same expected profits regardless of whether they enter in period 1 or 2).
4. If one recalculates the equilibrium for different demand curves, one can determine the effect of changes in uncertainty. As initial uncertainty increases, but holding constant expected demand, less entry occurs in period 1, more in period 2, and price fluctuations in period 1 are greater. This illustrates that the entry process and specifically the timing of entry will be driven by the value of the information one obtains from waiting. This value will depend upon what one learns about prices by waiting. If prices today are informative about prices tomorrow (as in this model), entry next period will be related to the current price level. If prices today are not informative about future prices (as in the model of

[15] This example is explained more fully in Illustration 1.

Illustration 1.

Equilibrium under uncertainty and sunk cost with time dependence

If Q is output determined at the beginning of period 1 before price is observed, then since demand is with equal probability either $100 - P$ or $90 - P$, equating supply to demand reveals that price will be either $100 - Q$ *or* $90 - Q$ with equal probability. In period 2, either someone will enter and drive price down to \$7 or no one will enter and price will remain at its period 1 level.

If a firm enters at the beginning of period 1, its expected profit, Π, is $\Pi = ½ (100 - Q) + ½ (90 - Q) + ½ \ \$7 + ½ (90 - Q) - \$10$ on the assumption (that is satisfied below) that $100 - Q > \$7$. Rewrite Π as $\Pi = 50 + 90 + 3.5 - 1.5 Q - 10$ or $133.5 - 1.5 Q$. In a competitive equilibrium with free entry, $\Pi = 0$.

Setting $\Pi = 0$ yields $Q = 89$, which implies from the demand curves in period 1 that price in period 1 is either \$11 or \$1. If price in period 1 is \$11, then entry of four firms occurs in period 2 in order to drive price in period 2 to \$7. If price in period 1 is \$1, no entry occurs in period 2 and price in period 2 remains at \$1.

the preceding section when demand uncertainty repeated itself each period and investments lasted only one period), entry next period is not related to current prices. Therefore, the entry process will depend upon the time dependence of the price structure which in turn depends on how uncertainty evolves and how long sunk investments last. This time dependence will be an equilibrium characteristic of the particular industry and will generally differ across industries. Hence, the entry process across industries will differ based on the fundamentals of underlying uncertainty and the nature of the irreversible investment.

5. Although not captured in the simple model, it is easy to see that an entrant should be concerned with both the cost and flexibility of his plant. An early entrant may be able to save money by choosing an inflexible plant, but this limits his ability later on to respond as industry conditions change. Therefore, if information about price volatility is revealed over time, later entrants will choose plants of different flexibility than early entrants.

There is no exit decision in the model, but it is easy to incorporate one. Suppose at the end of period 1, after prices are observed, that each firm already in the industry must decide whether to spend some amount, say \$2, to preserve its ability to produce in period 2. Clearly, any firm already in the industry will do so if price in period 1 exceeds \$2 because price in period 1 is, by assumption, a perfect predictor of price in period 2. But what if price in period 1 is low, say \$1. In that case, no firm would pay \$2 in order to sell at \$1 in the second (and final) period. Hence one will see exit of firms at the end of period 1. Enough firms must exit to drive that low price up to \$2 in period 2. The ability to exit is an option that is valuable to the firm because it reduces the firms' losses. If instead, the firm had to spend an extra \$2 as a sunk cost to enter initially and had no required ongoing expenditures of $2, the firm would be worse off because it would lose the ability to save the $2 by exiting at the end of period 1.

The exit decision will alter the number of firms in period 2, and so will affect price in period 2. By doing so, it affects the decision to enter in period 1! Entry and exit are different processes that affect each other in equilibrium and together determine the equilibrium price. By recognizing that an exit decision can cut off losses, the firm can decide how to structure its technology—high sunk costs with low operating costs or low sunk costs with high operating costs. A firm may opt for the latter even if it is more costly for any given production levels over a given time interval because it allows the firm to exit if times are bad and save some ongoing expenditures. For example, would a firm prefer to spend $10 up front to be able to produce one unit for each of two periods or would it prefer to spend $7 per period to be able to produce one unit per period? It depends. In the second scheme the firm may decide not to spend the $7 to produce in period 2 if it observes a low price in the first period.

Conversely, a firm may choose not to exit an industry if there is a possibility that next period (and subsequent periods) the price may be high. To see this clearly, consider a slightly more complicated model with at least three periods and a little bit more uncertainty. Suppose that at the end of period 1 if price is currently low (unlike in the simple model), there is a small chance that next period (and forever more) the price will be much higher. Well, the firm might pay to remain in the industry (i.e., not exit) in period 2 to see what happens in period 2. If price fails to be high in period 2, then the firm does not pay an additional amount to remain in the industry in period 3 and exits. But the point is that the expected price in period 2 may be low, yet the firm will still find it profitable to stick around, hoping to benefit (forever) from the high price which might materialize.

To better understand this, consider the following. Suppose the price in period 2 will be either $20 or $0 with probability ½ respectively. If it costs $15 to remain in the industry each period and this cost must be made before observing price, the expected profit in period 2 is negative ($10 – $15 = –$5), but that calculation ignores the fact that at the end of period 2 the firm will know whether prices in the future will be either $20 or $0. If the price is $20 forever, the firm earns $10 forever ($20 minus the $10 per period investment). If the price in period 2 is $0, the firm drops out at the end of period 2 and stops investing $10 per period. The important point is this: firms remain in an industry in period 2 even though the expected (period 2) profit is negative. They do so because they preserve an option to remain in the industry in the future should that prove to be profitable.

These admittedly simple and somewhat artificial models illustrate how complicated an analysis of entry and exit can be. It is not the case that the simple lower and upper bounds on price—determined as the minimum of the average variable cost and average cost curve respectively—correspond to trigger points for exit and entry in a dynamic world with uncertainty and sunk costs. Firms choose to stay in an industry when they appear to be losing money per period and will be hesitant to enter an industry even when price is currently high because price could subsequently fall. For example, United Airlines may choose to remain in the airline industry, rather than shut down and exit the industry, even though it is currently losing money because air fares may rise in the future, causing the firm to become profitable. The length of time an investment is sunk will influence these decisions. If no investments are sunk, or if there is no financial advantage to pay up front as a sunk cost rather than pay per period, the process of entry and exit is simple. There is one price in both the short and long run that will prevail—the price that corresponds to minimum average cost, C_2. But building in irreversibility of investments, uncertainty, and evolution of uncertainty over time (i.e., time dependence in price) greatly complicates the analysis. These last factors are fundamental determinants of industry equilibrium just as are underlying costs and demand.

The work of Dixit and Pindyk demonstrates how much the traditional analysis changes as a consequence of dynamics involving sunk cost and uncertainty.[16] Their models are much more complicated and realistic than the simple ones discussed here and rely on some very sophisticated techniques beyond the scope of this chapter. But their results and the intuition behind them are informative and illustrate several of the points in this chapter.

Imagine an industry where price today is informative about what price will be tomorrow. It does not perfectly predict future prices, as in the simple model discussed in this chapter, but instead gives an indication as to whether prices will be high or low next period. Also, suppose that to enter the industry a firm must make an irreversible investment that allows the firm to produce over some extended time period. Each period a firm in the industry must pay an amount to remain active next period; otherwise it exits the industry.

Dixit and Pindyck[17] develop such a model for the copper industry. If one were to use the simple model with no uncertainty, the values of C_1 and C_2 are approximately \$0.79 and \$0.88. That is, one would see exit when price hits \$0.79 and entry when price exceeds \$.88. Using a model incorporating the features of uncertainty and irreversible investment that was just discussed, they calculate that the actual exit and entry thresholds are \$0.55 and \$1.35. As these simple models suggested, firms will remain in an industry even when current profits are negative, and firms will not enter even when current profits are positive. Thus, the presence of uncertainty and long-lasting investments not only dramatically alters the timing and profitability of entry and exit but also changes the range of prices that can be observed in equilibrium. Moreover, one can use the model to calculate the likely price of copper. It turns out that copper producers will earn what look like supercompetitive profits about 60 percent of the time and will earn what look like below competitive profits about 30 percent of the time. Of course, expected profits over time equal zero as a condition of equilibrium. Stated differently, the copper price is outside the \$0.79 and \$0.88 range of the simple model about 90 percent of the time. The constraints that entry and exit place on price are much different in models that account for uncertainty and adjustment cost than in models that do not.

[16] DIXIT & PINDYCK, *supra* note 11.

[17] *Id.* at 264-67.

6. Empirical studies of entry and exit

Bain deserves credit for pioneering empirical studies on the reasons why some industries allow incumbents to earn excess rates of profit. Bain identifies whether certain industries have product differentiation, scale economies, large capital requirements, and cost advantage in order to understand why entry does not erode excess profits.

Although this chapter has questioned whether Bain's theoretical analysis allows him to claim that he has identified underlying structural features of the industry (i.e., ones that are exogenous), his studies do allow one to see what industry features coincide with high profits and lack of entry and are therefore valuable. A key issue is whether he has identified correlation or causation. If the conditions he identifies as barriers are endogenous (i.e., determined as a result of competition) then his observations, though interesting as industry descriptions, do not amount to a theory of how industry performance depends on entry barriers.

There is good deal of controversy as to whether there is empirical support for the claim that entry barriers lead to high profits.[18] There have been relatively few studies of the speed with which excess profits are eroded, but one common finding, that goes back at least to Stigler,[19] is that excess profits are eliminated very slowly in concentrated industries.

Sutton[20] is the most recent systematic study across industries in which intensive study of individual industries is used to identify underlying structural conditions to predict competitive performance. He does not focus on rates of return (which is what motivated Bain and Stigler) as much as on equilibrium concentration. His work emphasizes the distinction between exogenous and endogenous sunk costs and shows how endogenous sunk costs are an outcome of a (static) competitive game. For Sutton, the competitive game is taken as exogenous, but research and development and advertising may not be. Accordingly, it would not make sense to treat research and development or advertising as an exogenous barrier.

There have been several empirical studies of the entry and exit process.[21] Some of the key findings of those studies are:

1. There is an enormous amount of entry and exit of firms in manufacturing. About 40 percent of firms in an industry were not there five years earlier and will not be there in five years. Industries with lots of entry also have lots of exit.
2. Entrants and exiters are typically quite small relative to the industry average.
3. Entrants that have no experience in related industries are much smaller and fail more quickly than entrants with experience. For example, entrants with

[18] For a more extended discussion of some of these issues, see CARLTON & PERLOFF, *supra* note 6, at ch. 8.

[19] *See* GEORGE STIGLER, CAPITAL AND RATES OF RETURN IN MANUFACTURING INDUSTRIES (1963).

[20] *See* SUTTON, SUNK COSTS, *supra* note 4; SUTTON, TECHNOLOGY, *supra* note 4.

[21] *See* Timothy Dunne, Mark J. Roberts & Larry Samuelson, *Patterns of Firm Entry and Exit in U.S. Manufacturing Industries*, 19 RAND J. ECON. 495 (1988); STEVEN J. DAVIS, JOHN C. HALTIWANGER & SCOTT SCHUH, JOB CREATION AND DESTRUCTION (1996).

experience in related industries are about three times the size of entrants with no experience.

4. There are large differences across industries in both entry and exit rates.
5. For any one industry over time, the exit rates are more sensitive to business conditions than entry rates.
6. Industry expansion and contraction occurs largely through mature firms.
7. There is enormous heterogeneity in size across firms in the same industry, indicating, for example, different shut down costs and benefits.

Research on attacking the general problems of deriving equilibrium from exit and entry processes in a dynamic model with uncertainty and heterogeneity is just beginning. Early pioneering work did not focus on option values but did recognize and estimate heterogeneity amongst firms.[22] The theory and empirical methods of figuring out dynamic equilibrium are quite complicated. Recent work[23] has made significant progress in dealing with the theoretical and empirical problems arising in a model taking into account uncertainty, dynamics, and heterogeneity. But this body of research is just emerging and it is premature to draw general conclusions.

7. The use of entry barriers in antitrust and regulatory proceedings

Entry barriers are frequently an issue in antitrust cases and regulatory proceedings. Aside from the imprecision in its meaning, a problem with using the concept is that entry barriers are concerned with the long run, yet the long run may not be relevant for antitrust or regulatory proceedings. What often matters for antitrust and regulation is not what might happen in some year far off in the future, but what will actually happen now and in the relatively near future. Rather than focusing on whether an entry barrier exists according to some definition, analysts should explain how the industry will behave over the next several years. That will force them to pay attention to uncertainty and adjustment costs, the importance of which are recognized by some.[24]

The 1992 *Horizontal Merger Guidelines* of the Department of Justice and Federal Trade Commission[25] do a good job of explaining that entry matters in merger analysis only when it is timely (e.g., within two years) and of sufficient magnitude to keep price

[22] *See, e.g.*, Timothy F. Bresnahan & Peter C. Reiss, *Entry and Competition in Concentrated Markets*, 99 J. POL. ECON 977 (1991).

[23] *See, e.g.*, Richard Ericson & Ariel Pakes, *Markov-Perfect Industry Dynamics: A Framework for Empirical Work*, 62 REV. ECON. STUD. 53 (1995); Ariel Pakes, Michael Ostrovsky & Steve Berry, Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry/Exit Examples), (Harvard Institute of Economic Research, Discussion Paper No. 2036, 2005); P. Bajari, L. Benkard & J. Levin, *Estimating Dynamic Models of Imperfect Competition*, 75 ECONOMETRICA 1331 (2006).

[24] *See, e.g.*, Pindyck, *supra* note 11; RICHARD POSNER, ANTITRUST LAW: AN ECONOMIC PERSPECTIVE (2d ed. 2001); Richard Schmalensee, *Sunk Costs and Antitrust Barriers to Entry*, 94 AM. ECON. REV. 471 (2004); Preston McAfee, Hugo M. Mialon & Michael A. Williams, *What Is a Barrier to Entry?*, 94 AM. ECON. REV. 461 (2004).

[25] U.S. DEP'T OF JUSTICE & FEDERAL TRADE COMM'N, HORIZONTAL MERGER GUIDELINES § 0.1 (1992) (with Apr. 8, 1997 revisions to Section 4 on efficiencies), *reprinted in* 4 Trade Reg. Rep. (CCH) ¶ 13,104.

from rising above current levels. One may quibble about the "timely" definition, but the point is clear. What should matter to policymakers is how fast entry erodes any price increase caused by a merger and not whether it eventually does so. In litigated cases, especially monopolization cases under Section 2 of the Sherman Act, emphasis should be placed on how long entry will take before it erodes any temporary market power created by some challenged practice and whether the practice creates efficiency benefits. The rule of reason is the correct approach here, in which the costs and benefits are compared, but it often seems that possible efficiencies do not always receive full consideration. There seems to be a negative connotation to the word entry barrier and no recognition that, without some entry barrier, there may be no incentive to create new products or services. Indeed, as Demsetz has observed, property rights could be defined as the ultimate barrier to entry.[26]

In some regulatory proceedings (e.g., telecommunications and railroads), there has been a tendency to rely on contestability theory as a guide to setting price. Contestability theory is often described as a theory in which there are no barriers to entry or exit, so that instantaneous entry or exit is possible.[27] This chapter has explained why it is a mistake to confuse barriers to entry with factors affecting the timing of entry. They are two distinct concepts. But contestability theory, as commonly implemented, ignores uncertainty and adjustment costs.[28] As has already been shown, the ability of entry and exit to constrain price can differ depending on the presence of uncertainty and adjustment costs. Therefore, the equilibrium in the absence of uncertainty and adjustment costs need not be the same as the equilibrium with uncertainty and adjustment costs, especially for growing industries. Where the two equilibria differ, one obtains misleading policy advice by ignoring uncertainty and adjustment costs.

8. Conclusion

The words that one uses often can have unintended consequences when their meaning is unclear or even when their meaning is clear to the speaker but not to the listener. Barriers to entry as identified by Bain, is a confusing concept. Barriers to entry as defined by Stigler is clear, but perhaps strange, because the words mean something other than what would naturally come to mind. In any case, the failure of the concept of barriers to entry to incorporate a time dimension means that it is a concept in need of additional embellishment in order to be useful in antitrust or regulatory proceedings.

Putting aside the precise definition for barrier to entry, the idea that entry or exit can keep price in some narrow price band is wrong for certain industries. Underlying uncertainty and sunk costs can significantly alter the equilibrium from one based on

[26] Harold Demsetz, *Barriers to Entry*, 72 AM. ECON. REV. 47 (1982).

[27] Such a description fails to recognize adequately the contribution of this literature to one's understanding of sunk costs. *See* BAUMOL ET AL., *supra* note 9.

[28] Martin Weitzman proved that in continuous time contestability theory simplifies to a theory of constant returns to scale, that is, a model with no adjustment costs. *See* Martin Weitzman, *Contestable Markets: An Uprising in the Theory of Industrial Structure: Comment*, 73 AM. ECON. REV. 486 (1983).

models that ignore these factors such as models based on contestability theory. Entry and exit are more complicated phenomena than these models suggest. Using such models to formulate antitrust or regulatory policy can be a mistake.