Sunk Costs and Market Structure

Price Competition, Advertising, and the Evolution of Concentration

John Sutton

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Foreword

The literature on industrial organization has undergone a radical change over the past decade. This change has involved the reformulation of many traditional arguments within the subject in terms of (explicitly game-theoretic) oligopoly models. In developing this program, researchers have rediscovered an old difficulty. There are usually many ways of designing a game-theoretic model, which appear equally reasonable a priori. Moreover, within any particular model there are often many outcomes that can be supported as equilibria. This richness in modeling has made it much easier to provide a theoretical rationale for a wide range of observed phenomena: from predatory pricing to vertical restraints, our tool kit has been greatly enriched. The sting in the tail, however, lies in the old taunt, "With oligopoly, anything can happen." In explaining everything, have we explained nothing? What do the theories exclude?

Put more constructively, if the results of game-theoretic analyses depend delicately on a range of factors that are impossible to identify or proxy empirically, then how can we implement and test such theories? The currently popular response is to focus on some particular industry for which we can tailor-make a specific oligopoly model. By relying on arguments specific to this industry, it may be possible to restrict the class of admissible specifications quite tightly a priori and so generate a range of testable predictions. This approach has proved to be quite fruitful over the past few years, and has provided the beginnings of a response to the above line of criticism. At the same time, an increasingly skeptical attitude has been evident as to the usefulness of the kind of cross-industry studies that formed the staple of the traditional literature.

In this book I attempt to develop a different, though complementary, approach to these issues. I draw out, within a general theoretical framework, some fairly robust results that hold across a broad range of model specifications. These properties do not depend delicately on such nonmeasurable features of the model as might be expected to vary substantially from one industry to another. They can therefore provide a framework within which we can analyze a relatively broad class of industries and provide a new foundation for the kind of cross-industry comparisons of structure that became popular in the earlier literature, following Bain's pioneering work. Thereafter, by exploring the experience of each industry, or group of industries, in some detail, it is possible to explore various special cases arising within this general theoretical framework.

One of the most striking features of the industrial organization field generally is the cleavage between the traditional empirical literature on the one hand and the recent game-theoretic literature, with its associated single industry empirical studies, on the other. This book constitutes a first step toward building some bridges between certain recent theoretical advances and one major part of the traditional agenda of the subject.

1

An Introductory Overview

1.1 A Statistical Regularity

Many authors have observed that the ranking of industries by concentration level tends to be closely similar from one country to another: an industry that is dominated by a handful of firms in one country is likely to be dominated by a handful of firms elsewhere too.[1] This "statistical regularity" has occasioned a wide range of response in the literature. The large majority of studies argue in favor of the existence of such a regularity and interpret it as a reflection of the fact that the pattern of technology and tastes that characterize a given market may be expected to be similar across different countries. For this reason, the industry's equilibrium structure may in turn be similar from one country to another. While some authors have regarded this similarity of structure as rather trivial and of little interest, many, if not most, authors have seen it as providing considerable encouragement for the view that the underlying pattern of technology

1. The issue was raised by Bain (1966). An ambitious early study by Pryor (1972) covered twelve countries. That study suffered from one serious limitation, however, insofar as it used official statistics on concentration ratios that involved different levels of aggregation for certain industry groups, in different countries. Phlips (1971) avoided this problem by taking advantage of the newly available statistics for EEC countries, which were based on a common set of industry definitions. A similar approach was followed by George and Ward (1975). More recently, Connor et al. (1985) have carried out comparisons for industries within the food and drink sector over OECD countries. All these authors conclude that a high degree of correlation exists, whether comparisons are made on a pairwise basis (regressing U.K. concentration levels on U.S. levels, etc.) or otherwise. Many studies indicate an unusually wide disparity of experience between the United Kingdom and the United States; the relatively poor correlation obtaining in this case was emphasized by the early study of Shepherd (1961).

and tastes strongly constrains equilibrium structure (Scherer 1980, Caves 1989, Connor et al 1985).

Closely related to this empirical regularity is an important strand in the traditional literature on industrial structure, which aims to explain differences in concentration across industries by reference to a small number of candidate explanatory variables that are taken to reflect basic industry characteristics. Typically, the degree of scale economies, the intensity of advertising, and the level of R&D expenditure have been regarded as key variables. This study attempts to develop a new approach to this issue.

Most of the existing empirical work on cross-sectional differences in structure is based on an appeal to the structure/conduct/performance paradigm of Bain (1956). Within that paradigm, it is supposed that a one-way chain of causation runs from *structure* (the level of concentration) to *conduct* (the degree of collusion), and from conduct to *performance* (profitability). Structure, in this setting, is explained by the presence of certain barriers to entry, whose height can be measured by the degree of scale economies in the industry and by observed levels of advertising and R&D outlays relative to industry sales. This approach has for example motivated various empirical studies that seek to explain structure by regressing observed concentration levels on measures of scale economies, advertising intensity, R&D intensity, and so on (see chapter 5).

Bain's pioneering studies laid the foundations for a generation of subsequent work. His approach, however, has been subjected to considerable criticism both by empirical researchers during the 1960s and 1970s and by contributors to the game-theoretic literature of the past decade. The approach taken in this study, like many contributions to the recent game-theoretic literature, differs sharply from that of the Bain paradigm. The key theoretical differences between the two approaches are set out in section 1.4.

This volume also differs from much of the earlier literature in terms of its empirical focus. Here the starting point of the analysis lies in an alleged tendency for a given industry to be less concentrated, in those countries in which the size of the market (as measured by the total volume of sales) is larger. This negative relationship between market size and concentration has been noted by several authors (Phlips 1971, George and Ward 1975, Schmalensee 1989), but it has not received much attention in the past. After all, it seems a "natural" result and one that can be immediately explained by reference to traditional ideas. Put loosely, we might expect that given any particular con-

figuration of barriers to entry, an expansion in the size of the market will raise the profitability of incumbents and so induce more potential entrants to surmount that barrier, thus leading to a fall in concentration.

This relationship between market size and market structure stands at the center of my analysis, which is developed in three steps. The first step lies in describing a new theoretical rationale for the appearance of a negative relationship between market size and concentration, which is quite different to that embodied in the post-Bain literature. The second step involves the central claim of the present study. It says that this size-structure relationship, which has traditionally been seen as holding across industries generally, is in fact only valid for a certain group of industries. Most important, it is not valid for those industries in which advertising and R&D outlays play a significant role. In this latter context, it is argued that the negative relationship between market size and concentration levels breaks down for reasons that are quite fundamental.

The third step in the argument lies in developing some implications. Once the mechanisms leading both to the appearance of this negative relationship and to its breakdown in advertising and R&D intensive industries is understood, we are led to a new way of organizing many long-standing ideas regarding the determinants of cross-industry differences in structure. Those few empirical relationships that have emerged consistently in earlier empirical work now emerge as corollaries of certain relationships implied by the present theory. At the same time, we are led to a series of new approaches to a wide range of phenomena, noted in earlier empirical work, and to some new insights regarding a number of long-standing controversies.

But why should the relationship between market size and market structure merit such attention in the first place? The reason for this lies in the fact that the present characterization of the size-structure relationship represents one of the relatively few robust theoretical results to have emerged from the recent game-theoretic literature. To see the point of this remark, it is necessary to digress a little.

1.2 Game-Theoretic Models

Within the recent game-theoretic literature, numerous authors have sought to examine the long-run issues surrounding the determination of industrial structure (see Dasgupta and Stiglitz 1980, Shaked and

Sutton 1982, and Vickers 1986). One feature basic to this game-theoretic literature, however, is that the results of such analyses tend to depend delicately on the precise form of the underlying game.

Game-theoretic oligopoly models employ various simple building blocks that carry key distinctions of empirical interest. For example, we may capture the notion of the toughness of price competition by distinguishing a "Bertrand" formulation, a "Cournot" formulation, or a "joint profit maximization" formulation. Again, we may capture the presence or absence of some strategic asymmetry in the firms' relations to each other by contrasting a sequential moves formulation with a simultaneous moves formulation.

But these distinctions can often be mapped into empirical categories only in a rather loose and informal way. We may be willing to accept some particular formulation as a reasonable representation for some specific market, at least in the sense of a "prior" or null hypothesis. But if we aim to investigate statistical regularities that are presumed to hold across a range of different industries, between which the toughness of price competition or the degree of strategic asymmetry may vary, it may be extremely problematic to identify any measurable market characteristic that can act as an adequate proxy in capturing such distinctions.

Many researchers have come to feel that a natural response to such difficulties is to focus analysis on some specific market, or some set of virtually identical markets, so that we can tailor-make the oligopoly model to fit that specific context. The "ultra-micro" work to which we are led along this route is now one of the most lively areas of empirical research in industrial economics.[2]

These observations have led to a growing skepticism about the value of searching for statistical regularities that hold across a broad run of different industries. After all, if current theory indicates that most results are delicately dependent upon certain factors that are liable to vary widely across different industries and we cannot measure or proxy these in any satisfactory way, then the basis of running cross-industry regressions might appear to be somewhat dubious.

A central thesis of the present study is that this currently popular view is unduly pessimistic. Moreover, a too-rigid adherence to such

2. For a selection of such studies, see the *Journal of Industrial Economics* symposium of June 1987. Influential contributions include Hendricks and Porter (1988) on the auctioning of offshore oil leases, Slade (1987) on gasoline price wars, and Bresnahan and Reiss (1990) on the relationship between the size of towns and the number of retail outlets.

a view runs the risk of abandoning a central part of the traditional agenda of the subject, which concerns the investigation of regularities of behavior that hold across the general run of industries.

The point of departure lies in the observation that a fundamental trade-off exists between the degree of precision of predictions that obtain across a class of models and the breadth of applications of that class. Tight predictions may demand quite stringent a priori restrictions on the model(s); but it may nonetheless be possible to find some (necessarily weak) predictions that are robust in the sense that they hold across a wide class of models—and so lend themselves to implementation across a correspondingly broad set of different industries.

The approach taken in what follows, then, is to begin by looking to such robust predictions as will hold across a wide class of reasonable models, and to use these predictions as a basis for cross-industry regressions. What kind of results can be obtained at this level of generality? It turns out that certain robust results can be obtained that relate to the specification of a *lower bound* to the equilibrium level of concentration as a function of the size of the market. The properties of such bounds are studied in detail, and their role is examined empirically by reference to cross-country comparisons of industry structure. A number of ancillary results are also obtained at this general level, which again can be investigated by reference to cross-country comparisons.

In parallel with the development of this general framework, various special cases within the theory are investigated, and these may be used as a vehicle both for tracing the evolution of particular industries and for generating a richer menu of testable predictions appropriate to a correspondingly narrower domain. The next three sections are devoted to providing a chapter-by-chapter summary of the study as a whole.

1.3 An Outline of the Theory

The theory developed below derives from the recent vertical product differentiation literature. The starting point of this analysis, as developed by Shaked and Sutton (1982, 1987) and elaborated in Sutton 1989a, lies in the observation that advertising and R & D can both be thought of as sunk costs incurred with a view to enhancing con-

sumers' willingness-to-pay for the firm's product(s). Focusing attention on this relationship makes possible a simple unified treatment of these two contributory factors. Now R & D and advertising outlays are choice variables to the firms, and so their levels must be determined endogenously as part of the specification of industry equilibrium. The role of scale economies, on the other hand, can be introduced by treating the acquisition of a single plant of minimum efficient scale as involving an element of sunk cost that must be incurred by all entrants, and whose level is determined exogenously by the nature of the underlying technology.

The central focus of the present theory lies in unraveling the way in which these *exogenous* and *endogenous* elements of sunk cost interact with each other in determining the equilibrium pattern of industrial structure. This theoretical framework is set out in detail in chapters 2 and 3. In what follows, a brief description of the main features of the theory is presented, as a prelude to summarizing the contents of later chapters.

Exogenous Sunk Costs (i)

The way in which the central notion of *sunk costs* is captured in the present study is by modeling industry equilibrium in terms of a two-stage game. At stage 1 of the game firms incur fixed outlays, which are associated with acquiring a single plant of minimum efficient scale (setup costs), and developing and establishing a product line (possibly incurring advertising and R & D outlays). These fixed outlays incurred at stage 1 of the game are treated as sunk costs in analyzing price competition at the second stage of the game. In the latter stage of the game, all firms are assumed to operate at the same constant level of marginal cost.

Consider first the case in which the only sunk costs involved are the exogenously given setup costs. Within this case it is useful to distinguish two subcases. The first subcase is that in which the various firms produce a homogeneous product. In this setting, as the size of the market (measured by the population of consumers) increases, the equilibrium number of firms entering the market increases, and so concentration declines indefinitely. To see this, note that entry occurs up to the point at which the (stage 2) profits of the last entrant cover the sunk cost incurred on entry at stage 1. But, for any given level of concentration in the industry, any increase in the size of the market

will tend to raise these profits and so induce further entry. Thus concentration declines indefinitely as market size increases (except under very special circumstances; see chapter 2).

This case, then, corresponds to some familiar limit theorems of the standard theoretical literature; and it offers one way of characterizing the traditional idea that scale economies become unimportant as a constraint on equilibrium structure in large economies. What may be less obvious is the way in which this process is affected by the nature of price competition at stage 2 of the game. When analyzing stage 2, concentration is taken as fixed—being inherited as a result of decisions made at stage 1, which are now irreversible (i.e., they embody sunk costs). Now this means that we can properly build into our analysis of this stage of the game (i.e., the stage 2 subgame) the traditional Bain hypothesis on conduct: that prices (and so unit margins) decline as concentration falls. Within the present theory, this notion is embodied in the form of a *function* linking concentration to prices or unit margins. This *function* will be affected by such features of the market as the physical nature of the product (homogeneous versus differentiated products) and the climate of competition policy (a strict or acquiescent approach to price coordination by firms). In what follows, references to the "toughness of price competition" in a market will always refer to this *function*—and *not* to the level of prices or unit margins observed at equilibrium. (In other words, differences in the toughness of price competition across two different markets relate to the way in which margins *would* differ between those markets were concentration held at the same arbitrary level in both.)

It is shown in chapter 2 that according as price competition is tougher in this sense, the equilibrium level of concentration will be correspondingly *higher*. The intuition underlying this result is simply that the anticipation of a tougher competitive regime makes entry less attractive, thus raising equilibrium concentration levels. One of the main attractions of the two-stage game formulation is that it allows a neat unraveling of this latter effect from the traditional Bain effects (higher concentration implies higher margins, and higher profitability). To sum up: where sunk costs are exogenous, and where firms offer a homogeneous product, the equilibrium level of concentration declines with the ratio of market size to setup cost and rises with the toughness of price competition.

Exogenous Sunk Costs (ii)

The next subcase to be considered is that in which firms offer products that are differentiated, but in which sunk costs are still exogenously determined. This case has been widely explored in the horizontal product differentiation literature; the archetypal example arises in simple locational models of the Hotelling kind (a brief description of these standard models is provided in chapter 2). In these models, consumers are spread over some geographic region, and they incur (psychic or transport) costs in purchasing from distant suppliers. Each firm may establish any number of plants, incurring a given setup cost per plant. Consumers thereafter make their purchases from the lowest-cost supplier, where the cost to the consumer consists of the price paid to the firm plus a transport cost that increases with his distance from the supplier.

In models of this kind, multiple equilibria are endemic. In general, for any given market size, we may find fragmented equilibria, in which a large number of firms each sell at one location, and concentrated equilibria, in which a small number of firms each sell at many locations. In chapter 2 the factors underlying the appearance of these two types of equilibria are set out; these factors are likely to vary widely from one industry to another. Thus, if we are interested in finding properties robust enough to be of interest in cross-industry studies, we cannot constrain possible equilibrium configurations here, beyond saying that the bound corresponding to the most fragmented configuration (single-product firms) forms a *lower bound* to equilibrium concentration. This bound declines with market size in the manner of the schedule described above for the homogeneous product case.

The case, then, is that in which the present theory *least* constrains the data; and so this set of industries provides the first of several illustrations of the inherent limitations of this theory, insofar as robust predictions appropriate to a broad cross-section of industries are involved. It is perhaps worth remarking, therefore, that even in this case the theory *does* in fact yield some quite sharp predictions. But these predictions depend in all cases upon market features that are likely to vary widely from one industry to another and that are difficult to measure or proxy in many instances. (Shaked and Sutton 1990).

Bringing together these remarks on the two subcases, then, a central conclusion for the *exogenous sunk cost* regime may be phrased

as follows: an increase in the size of the market relative to setup costs may lead to the appearance of indefinitely low levels of concentration in these industries. It is precisely this property that breaks down once we turn to the next case.

Endogenous Sunk Costs

We now turn to the case of *endogenous sunk costs*. These cost components may be of various kinds; the two most obvious examples, though not the only ones, are advertising and R & D outlays (see chapter 14). Suppose that, by incurring greater advertising (or R & D) outlays at stage 1 of the game, a firm can enhance the demand for its product at stage 2 (i.e., for any prices set by other firms, the demand schedule of the firm in question shifts outward). Then it is fairly obvious that the game played at stage 1 might involve a competitive escalation of outlays by firms and so lead to higher sunk costs being incurred at equilibrium. It is also fairly obvious that the larger the size of the market—and so the profits achievable at stage 2—the greater might be the sunk costs thereby incurred at equilibrium.

What is not obvious is that this is not merely a *possible* outcome; rather, on examining a range of different oligopoly models, an unusually robust result arises in this case, which runs contrary to that found in the exogenous sunk cost case. This result says that under very general conditions a lower bound exists to the equilibrium level of concentration in the industry, no matter how large the market becomes.

The level of this lower bound depends on the degree of demand responsiveness faced by an individual firm to increases in its fixed (advertising or R & D) outlays at stage 1 of the game. The higher the degree of responsiveness, the higher will be the lower bound to equilibrium concentration levels in the industry.

The central assumption of this study, then, is that across a certain range of industries, advertising works; loosely stated, it is postulated that the degree of responsiveness of demand to advertising outlays for any one of a number of competing firms always exceeds some minimal level. An exact statement of this assumption must be deferred to chapter 3.

Under these circumstances, increases in market size cannot lead to a fragmented market structure as the size of the market increases. Rather, a competitive escalation in outlays at stage 1 of the game

raises the equilibrium level of sunk costs incurred by incumbent firms in step with increases in the size of the market—thus offsetting the tendency toward fragmentation.

The importance of this simple but basic result lies in the fact that it holds over an extremely wide class of oligopoly models. For example, the result holds independently of whether each firm offers a single product or a range of products. It holds independently of the form price competition takes at stage 2 of the game (Bertrand, Cournot, etc.). Furthermore, it is not affected by altering the sequence of moves in the entry stage of the game (simultaneous entry, sequential entry, etc.). The degree of robustness of this result to changes in model specification makes it a suitable candidate for investigation in a cross-industry setting.

The above comments on the *exogenous* sunk costs case imply that, if we confine attention to some set of industries in which advertising and R & D outlays are insignificant and examine how concentration varies with the size of the market across different countries, then we should expect the lower bound to observed concentration levels to fall as the size of the market rises relative to the setup costs incurred in entering the industry. The central prediction of the theory is that this relationship should break down among advertising-intensive industries. The precise way in which the relationship fails, and the testable implications of this result, are developed in chapter 3.

Further themes explored in chapter 3 include the question of how exogenous sunk costs *interact* with endogenous sunk costs in determining industrial structure. Attention is also directed toward various special cases arising within this general theoretical framework. The most important of these special cases relates to the role played by "first-mover advantages" in determining equilibrium structure and to the factors leading to the evolution of dual structure—in which a small number of leading firms spending heavily on advertising and enjoying large market shares coexists with a possibly large fringe of nonadvertisers who sell on price.

1.4 Econometric Tests

The theoretical framework developed in chapters 2 and 3, then, leads to a basic prediction about the way in which the market size/market structure relationship will vary between industries in which sunk

costs are exogenously given and those in which endogenous sunk costs such as advertising or R & D play a significant role. This book is concerned with exploring this and other predictions of the theory within the context of a group of advertising-intensive industries.[3] In chapter 4, the rationale underlying the selection of this group of industries is set out in some detail. Broadly, the aim was to find a group of cognate industries in which R & D played an insignificant role and in which levels of advertising intensity were high on average, but varied widely across different industries within the group. Based on these criteria, the food and drink sector provided an obvious choice. This sector, which comprises about one-eighth of all manufacturing industry in the countries studied, has both the highest level of advertising intensity among all two-digit SIC groups and is among the lowest of all such groups in terms of R&D intensity.[4] This study is based on the experience of twenty narrowly defined food and drink industries across six countries (France, Germany, Italy, Japan, the United Kingdom, and the United States). These industries divide into two groups. In the first group, advertising outlays are extremely low in almost all cases; and these industries provide a benchmark case corresponding to the exogenous sunk cost case of the theory. In the other group of industries, the levels of advertising intensity are moderate to high; and the evolution of this group is examined by reference to the endogenous sunk cost case.

In chapter 5, a cross-sectional econometric analysis of observed concentration levels is presented, the results of which are consistent with the theory. These results, moreover, are *not* consistent with the

3. There are good reasons to divide the task of implementing the theory in this way, in spite of its emphasis on the similarities of the advertising and R & D cases. In the case of advertising-intensive industries, a great deal can be learned from cross-country comparisons of structure, since the sunk costs incurred by a firm in advertising its product in one country do not carry over to other countries. (Its brand image must be established anew in each country.) This is not so for R & D outlays, and most R & D-intensive industries are best treated as unified global markets. It is also worth remarking at this point that even at the theoretical level, the case of R & D is somewhat more complex than that of advertising, and an adequate treatment of the R & D case requires some extensions of the theoretical framework developed here (see chapter 14).

4. In the United States, for example, the food and drink sector accounted for 12% of the total value of production in manufacturing in 1980 (Connor et al. 1985). The level of private R&D expenditure as a proportion of sales was equal lowest with textiles and apparel, at 0.4% in 1975 (Scherer 1980, p. 410). The level of advertising expenditure relative to sales far outruns that of any other sector. Food and tobacco advertising accounted for 32% of advertising of all manufactured products in the United States in 1979, but for only 12% of all sales of manufactures (Connor et al. 1985, p. 80).

alternative view, that observed advertising levels can be regarded as exogenously given, that is, determined by product characteristics and other factors, independent of market size.

A second important theme developed in chapter 5 is that those few statistical regularities that have emerged more or less consistently in the earlier literature in this area can be shown to follow as a *consequence* of the basic regularity identified here. Thus the present theory, as well as generating new findings, appears successfully to provide an explanation consistent with these well-known empirical relationships.

Cross-industry regression results always invite alternative interpretations, however, and in the following industry studies an attempt is made to probe the validity of the interpretation offered here by investigating whether the pattern of evolution of structure in these two industry groups exhibits those different qualitative features implied by the theory.

1.5 Industry Studies

The core of this book consists of a matrix of industry studies, which has been compiled using a combination of published market research reports and a lengthy program of meetings with senior marketing executives in the industries concerned. An attempt was made, whenever possible, to provide accounts of the industry that were complete relative to the theory. One aim of this exercise is to provide the appropriate background information to readers who may wish to explore alternative explanations for the statistical regularities of chapter 5.

By building up a detailed profile of each industry, it was possible to go much further in probing the validity of the theory than would have been possible solely on the basis of a cross-sectional econometric analysis. Apart from directly testing the implications of the theory, moreover, these industry studies make possible a number of ancillary exercises. The presentation of these studies has, for expositional reasons, been arranged around a number of major themes.

(a) Testing the Theory I: Two Mechanisms

The central idea of the theory lies in the claim that two qualitatively different mechanisms operate to prevent certain fragmented configu-

rations from persisting over time—the claim is that such configurations are not (Nash) equilibria, that is, they will be broken because, in such a configuration, it will always be optimal for one firm to deviate in a way that destroys that configuration.

(i) With the exogenous sunk costs model, a too fragmented configuration will break down because it is impossible to maintain price-cost margins sufficient to generate a normal rate of return on the setup costs incurred in establishing plants; and while this is perfectly possible in the *short run*, it is not consistent with a *long-run* equilibrium situation in which obsolete plants need to be replaced periodically. Attempts by firms to coordinate prices to a degree sufficient to recover these outlays will fail unless a level of concentration is achieved that exceeds the lower bound consistent with equilibrium (see chapter 2 for details).

(ii) Within advertising-intensive industries, a second mechanism operates to exclude the persistence of certain fragmented configurations. In this setting, as has already been noted, the mechanism involves a competitive escalation of advertising outlays in the initially fragmented industry.

The first theme explored in the industry studies lies in examining whether the histories of these two groups of industries provide any evidence for the alleged operation of these two distinct mechanisms. Although these mechanisms recur throughout many of the chapters that follow, the contrast between them is best illustrated by reference to chapters 6 and 8.

Chapter 6 is devoted to the application of the exogenous sunk cost model to a study of the salt and sugar industries. These two industries are characterized by a high degree of product homogeneity and a fairly high level of setup cost relative to market size. The theory predicts that, in this setting, a process of free competition will lead to a highly concentrated structure. The experience of the salt industry offers a striking example of this mechanism. In the case of U.S. and British industries, whose history is relatively well documented, the market was initially quite fragmented. Both industries were characterized by strong price cutting, especially in periods of declining demand and repeated attempts to bring about price coordination among the many firms ended in failure. In each case, poor profitability led to a mixture of exit and of merger and acquisition activity, and it was this process which in turn led to a consolidated industry

structure within which unit margins were stabilized and profitability recovered. Current concentration levels in these industries greatly exceed those levels that have been considered "warranted" by previous observers (the traditional notion of warranted concentration levels is discussed in the next section).

The central novelty of the present theory in this context is that the equilibrium level of concentration is argued to depend inter alia on the toughness of price competition in the market. The theory predicts that if institutional factors cause price competition to become less tough (in the sense that higher unit margins can be sustained at any *given* level of concentration), then the *equilibrium* level of concentration will be correspondingly lower. The obvious way to probe the validity of this argument is to look to cases in which institutional factors impinge to a varying degree on the free play of competition in different markets. Except where state monopolies exist, the salt industry has in most cases operated with minimal intervention by the authorities. Policy measures in the industry have for the most part been "pro-competitive," in that they have involved no more than occasional attempts to limit or prohibit price fixing. But the sugar industry, in contrast, usually enjoys the support of a strong agricultural lobby, and the authorities' approach to price determination in the industry has varied widely, both across countries and over time. In some cases, the authorities have favored unfettered competition; in others they have effectively determined industry margins, with a view to either stabilizing or rationalizing industry structure. In others, policy has varied sharply in different periods.

In contrast to the salt industry, which is highly concentrated everywhere, the sugar industry shows a wide divergence of structure across countries. In chapter 6, it is argued that these differences in structure can be traced directly to the differences in policy regime in a manner consistent with the theory.

The cases of salt and sugar, then, illustrate the way in which the toughness of price competition impinges on the determination of structure in those industries where setup costs are high in relation to market size. The evolution of concentration in these two industries stands in sharp contrast to the pattern observed in advertising-intensive industries.

Chapter 8, a pivotal chapter, begins the exploration of advertising-intensive industries by examining the evolution of structure in the frozen food market. This case is of special interest, as this is one of the

rare examples of an advertising-intensive industry within the food and drink sector whose origin is both recent and sharply defined. The industry's beginnings can be traced to the development of certain freezing processes in the 1930s, and the early history of the industry, especially in the United Stated and the United Kingdom, is well documented.

The themes explored in chapter 8 provide a clear illustration of the mechanism postulated in the present theory as applying to industries that exhibit endogenous sunk costs. The setup costs incurred in entering the frozen food industry are quite low; and the market includes both a retail segment (within which advertising is quite effective) and nonretail segments in which buyers choose suppliers almost wholly on the basis of relative price. Under such circumstances, the theory predicts that increases in market size may lead to an indefinite expansion in the total number of firms, but that the (advertising-sensitive) retail sector will remain concentrated, while advertising outlays by leading sellers in the retail sector expand in step with the size of the market. (The theoretical basis for the evolution of this dual structure is described in chapter 3.) A competitive escalation of advertising outlays will necessarily lead to a situation in which only a small number of firms survive and dominate the retail segment of the market.

This process is well illustrated by the history of the frozen food industries of the United States and the United Kingdom. In each case, it is possible to pinpoint the exact phase at which firms became partitioned into two discrete groups: a high-advertising group selling primarily to the retail sector, and a nonadvertising group selling solely to the nonretail sector. Indeed, it is possible to trace the way in which firms situated between these two groups faced declining profitability until this split was achieved.

While the frozen food industries provide an unusually clear-cut illustration of this process, the same process can be seen to operate in a wide number of instances explored in later chapters. In many of these cases, however, the appearance of the high advertising group can be traced to the turn of the century, and documentation of the structure of the industries at that time is often sparse. Moreover, the partitioning of firms into two groups appears to have taken place in a more gradual and less dramatic fashion than in the case of the frozen food industry. This may in part reflect the lesser scale and effectiveness of advertising prior to the advent of television. These

qualifications apart, however, the same process appears to have operated across a wide range of those advertising-intensive industries that are explored in later chapters.

(b) Testing the Theory II: Comparative Static Predictions

A central issue concerns how the exogenous and endogenous elements of sunk costs *interact*. How does a higher level of setup costs affect equilibrium advertising outlays and equilibrium structure, other things being equal? Since other things are rarely equal, examining this question empirically poses considerable difficulties. As far as theoretical predictions are concerned, chapter 2 shows that a rise in setup cost will lead to a more concentrated equilibrium structure; this modest result, however, is the only robust comparative static result within the endogenous sunk cost model. In respect of advertising, a rise in setup costs from an initially very low level will at first imply a rise in the advertising-sales ratio. As setup costs continue to rise, however, the advertising-sales ratio may continue to rise, or it may fall. The outcome will depend delicately on the details of the model—and most importantly on the extent to which increases in total *industry* advertising affect total *industry* sales. This feature of the market will differ sharply across different industries, so that no robust result is available.

Investigating such comparative static properties is made difficult, in general, by the presence of a multitude of industry-specific characteristics whose effect may be extremely hard to quantify. The analysis of chapter 12 takes advantage of the fact that two pairs of industries within the present sample offer an unusually helpful context in which to examine this issue, for in each case the two industries in question have setup costs that differ by a very large factor, while the other economic characteristics of the industry are closely similar. The production of instant coffee involves setup costs very much greater than those incurred by the typical producer of ground (or roast and ground) coffee. Within the confectionery industry, the setup costs incurred by producers of mass-market chocolate confectionery items exceeds by an order of magnitude the costs incurred by the typical producer of sugar confectionery. In each of these cases, we find that the industry with the higher setup costs is more highly concentrated. Furthermore, differences in setup costs vary across different market segments; and a detailed analysis of how concentra-

tion differences mirror these differences in the pattern of market segmentation offers further support to this interpretation. Finally, it is shown that, within the confectionery sector, the high-setup cost (chocolate confectionery) industry displays a systematically higher advertising-sales ratio. This is not true within the coffee industry, however: here the advertising-sales ratio in the instant coffee market may be higher or lower than the ratio obtaining in the ground coffee market. While considerable caution is needed in interpreting these differences in experience, they appear in part to reflect the extent to which *total* advertising outlays for instant coffee are likely to expand total instant coffee sales (at the expense of total sales of ground coffee)—as the present theory implies.

(c) Special Cases: First-Mover Advantages

One important role played by the industry studies that follow lies in allowing an exploration of various special cases arising within the general theoretical framework. One such special case arises in respect of "first-mover advantages": to what extent do strategic asymmetries between early entrants to an industry and firms that enter later impinge on the equilibrium pattern of structure? Here, the results of a game-theoretic analysis suggest that the outcome depends delicately on the details of the model; and so in this context the focus is *not* on testing theory, but is merely exploratory. The aim is to examine on the basis of industry histories how such influences may play a part in accounting for the often wide divergence of structure in the same industry from one country to another.

It is often extremely difficult in practice, however, to decide whether or not a particular firm enjoys this kind of strong strategic asymmetry relative to its rivals. But there are some instances in which the historical and institutional background point to a sharp asymmetry between firms, of a kind that can reasonably be represented by a simple sequential entry model. Chapter 9 examines three industries in which a very clear-cut first-mover advantage was present and traces the apparent consequences of this strategic asymmetry for the evolution of structure.

In the prepared soups industry the contrast between the United Kingdom and the United States is of particular interest, as the same two firms dominate each of these markets. Their roles in the two markets are precisely reversed, however, with the U.S. market leader Campbell filling the same role in the United Kingdom market as does

the U.K. market leader Heinz in the U.S. market. This similarity of roles extends to such areas as the two firms' pricing policy, their decisions as to whether to supply retailers' own-label products, and so on; and these differences can be traced to the fact that each country's current market leader enjoyed a first-mover advantage over its rival in that country. In this case, the structure of the industry is closely similar in both countries, and only the roles of the leading players differ.

A second example illustrates how the presence or absence of first-mover advantages may exert an apparently profound effect on the overall structure of the industry. In the margarine market, Unilever enjoyed a strong first-mover advantage in the various European markets, where it still enjoys a leadership position. In the U.S. market, however, a strong agricultural lobby effectively stifled the development of the margarine market until the 1950s, by which time a number of major U.S. food producers had developed a position in the industry. Thus, no one firm had a first-mover advantage on the eve of the rapid takeoff in margarine sales, which coincided with the growth of television as the dominant advertising channel during the 1950s. The outcome was a less concentrated overall structure, in which Unilever competes on a more or less equal basis with two major indigenous food producers (Kraft and Procter & Gamble).

A third example of a first-mover advantage arises in the soft drink market. The Coca-Cola Company and Pepsico enjoy rough parity in the U.S. market, and the escalating competition between the two has played a central role in shaping the evolution of the U.S. industry. In Europe, however, a sharp asymmetry exists between the two, which can be traced to a series of events that took place during the Second World War. In return for its commitment to "put a Coke in the hand of every U.S. serviceman" for 10 cents, the Coca-Cola Company was made exempt from wartime sugar rationing, and the resulting ubiquity of Coca-Cola set the stage for the continuing dominance of Coca-Cola over Pepsi throughout postwar Europe. It is argued in chapter 9 that this asymmetry may be one of the contributory factors underlying the wide divergence in structure in this industry in the United States and Europe.

(d) The Limitations of the Theory

Another theme that runs through the industry studies relates to the limitations of the theory—indeed, one of the main virtues of con-

structing such a matrix of industry studies is that it provides an unusually detailed feel for both the strengths and the limitations of the theory. The limitations are of two kinds:

1. The theory predicts only a *lower bound* to equilibrium concentration levels. The lower this minimal equilibrium level, the less the theory constrains the data.

This implies an inherent limitation in the theory, the importance of which varies across different groups of industry. This issue is explored in chapters 7 and 8.

2. The usefulness of this theory rests on the assumption that the advertising response function depends only on certain (unspecified) product characteristics, which may be assumed similar across countries, and on *observable* institutional factors, which may differ across countries.

While this assumption appears, on the basis of the evidence, to be broadly reasonable, occasional instances arise in which idiosyncratic factors peculiar to certain firms or markets profoundly influence the extent to which advertising "works." A good illustration of this kind of effect, which highlights a potentially serious limitation to the value of theories of this kind, is provided by the experience of the mineral water market. Perrier's success in the French market, and its competitors' reaction to that success, transformed the structure of the industry within a few years. In no other country has Perrier, or any other company, achieved a similar success. Indeed, in the three main European markets (France, Germany, Italy), apparently closely similar market conditions coexist with levels of concentration that diverge very widely from one country to another (chapter 11).

(e) Controversial Cases

The final theme explored in the industry studies relates to the reexamination of some cases that have received heavy emphasis in the recent literature in terms of the present theoretical framework.

One of the most crucial U.S. antitrust cases in recent years was that brought by the Federal Trade Commission (FTC) against the Kellogg Company and its main rivals in the ready-to-eat (RTE) breakfast cereals industry. The case was interpreted in some quarters as constituting a test for a new departure in antitrust practice—a

departure in favor of government intervention in markets on the basis of an examination of structural features (concentration) per se, independent of any observed anticompetitive practices. Under these circumstances, it is not surprising that the case attracted an unusual degree of interest among industrial economists, and the causes of concentration in the industry became a much debated issue. The most popular theory to emerge was that proposed by Schmalensee (1978), who proposed that Kellogg's dominant position could be traced to a process of monopolization by product proliferation. By taking advantage of its status as a first-mover, Kellogg could allegedly fill all available product niches, leaving little scope for rivals to enter.

How does this view fit in with the present theory? In chapter 10, it is shown that the strategy of product proliferation can be seen as a special case of the general framework developed here; but placing it within this setting introduces as a additional and primary mechanism the same competitive escalation of advertising outlays that appears across the general run of advertising-intensive industries. It is argued that this difference in interpretation has possible implications for the effects of the type of remedy proposed by the FTC.

The evolution of concentration in the beer industry is the subject of the last of the industry studies presented in part II. Few industries have been more intensively studied; and particular attention has been focused on the U.S. industry, in which a steady rise in concentration over the past generation has led to a situation in which the emergence of a triopoly was becoming evident by the late 1980s.

The causes of this massive and sustained rise in concentration have been variously explained. Two lines of argument stand out in the economics literature. One emphasizes changes in brewing and bottling technology, which have led to a substantial rise in the minimum efficient scale of operation (m.e.s). Another view emphasizes the role of escalating advertising outlays by the major brewers and the consequent pressure on small and medium-sized firms. Some authors take an eclectic view, arguing that both of these factors have contributed to the changes in industry structure.

The central theme of chapter 13 is that within the context of the present theory, these two contributory factors, while both relevant, are not independent. Rather, they both play a part within the same unified mechanism that underlies the evolution of all of the advertising-intensive industries studied here. Under this interpretation, the exogenous changes in technology that have raised the mini-

mum efficient scale—and so the level of setup costs—have thereby stimulated and fueled the concomitant increases in advertising outlays, thereby accentuating what would otherwise have been a comparatively modest trend toward higher levels of concentration.

A central implication of this analysis relates to the likely effect of recent changes in technology, which appear to favor the relative efficiency of smaller breweries. It is argued in chapter 13 that such a reversal of earlier trends toward higher m.e.s. levels will *not* tend to lead to a dilution in industry concentration.

1.6 The Bain Paradigm Revisited

It may be of interest to readers familiar with the earlier literature to see how the ideas set out above relate to the Bain paradigm. Two points of difference are obvious. The first point relates to difficulties arising with the chain of causation story. As noted earlier, the Bain paradigm posits a one-way chain of causation running from structure to conduct to performance. The determination of structure, in this setting, is explained by reference to various barriers to entry, whose nature is taken to be exogenously given. It is not only within the recent game-theoretic literature that one finds the adequacy of such a story challenged: various early empirical studies drew attention to the need to consider a possible reverse link from conduct or performance to structure; but this was often presented as an econometric issue, in the sense that it might be dealt with by writing down a simultaneous equations system linking structure, conduct, and performance. This line of response did not prove very fruitful, however (see Schmalensee 1989 for a review). Within the present framework, the problem is tackled by looking rather to a reformulation of the basic theoretical model.The key contribution of the two-stage game formulation is that it offers a neat way of unraveling the two-way link between structure and conduct.

The second point of difference with the Bain paradigm lies in its treatment of observed advertising levels as a barrier to entry, whose level might be appealed to in explaining industry structure . This view of advertising as a barrier to entry was taken to its logical conclusion in a series of studies that regressed observed levels of (or changes in) concentration on a series of explanatory variables, including the level of scale economies, the level of advertising intensity, and

so on. (See chapter 5 for a comparison of the results of such studies to this study.)

Within the present approach, observed advertising levels are endogenously determined. This difference is of fundamental importance, and it raises some serious questions about the way in which the barriers to entry concept is commonly used. The role of this concept in the post-Bain literature lay in providing a rationale for why apparently high profits could be consistent with an absence of entry. If the height of such a barrier is endogenously determined, however, then considerable care is needed in formulating explanations of why observed concentration levels are high, or why high measured profit levels do not seem to stimulate new entry in certain industries.

Beyond these two rather obvious differences, however, there are a number of further ways in which the present approach departs from the traditional literature. One such feature relates to the relationship between product differentiation and advertising. In the traditional literature, these were often closely identified; advertising intensity was in fact widely used as a proxy for the degree of product differentiation. Such an approach is highly dubious on purely empirical grounds: many industrial sectors (such as engineering) involve both a high degree of product differentiation and minimal advertising. Within this study, a sharp analytical distinction exists between the effects of product differentiation per se and the effects of advertising outlays. In making this distinction, the concepts of horizontal and vertical product differentiation outlined above turn out to be extremely useful (section 3.5).

There is one final difference between the present approach and the earlier empirical literature that deserves a relatively full remark. This relates to the way in which a lower bound to industry structure was introduced by some contributors to the traditional literature, in the guise of a warranted level of concentration . The idea ran as follows: Consider an industry in which advertising and R&D play no role, so that a firm's costs may be identified with the (fixed and variable) costs of production alone.

Imagine that each firm operates subject to an average cost schedule that declines up to some critical level of output (the minimum efficient scale (m.e.s.) of operation) and is flat after that point. Any firm operating at a level of output below this point will suffer a competitive disadvantage, and at equilibrium the industry will be populated by

a number of firms operating at a level of output at or above the m.e.s. Thus a lower bound to industry concentration exists, corresponding to a configuration in which the industry consists of a number of firms operating at a level of output *equal to* the m.e.s. level. The associated level of concentration is referred to as the warranted level of concentration throughout much of the empirical literature of the 1960s and 1970s (see Anderson et al. 1975).

It was noted by many observers, however, that actual concentration levels seem to lie far above the levels warranted on the basis of such arguments (see Scherer 1980, and for a specific example, Anderson et al. 1975), whether the m.e.s. level used in calculating such bounds came from estimates of median plant size, or from independent engineering studies (see chapter 4).

How does this relate to the present approach? A strict formulation of the preceding argument depends necessarily on the assumption that the average cost schedule becomes perfectly flat once the m.e.s. level of output is exceeded. In applying the idea empirically, however, it was (and remains) customary to adopt an arbitrary definition of m.e.s. as corresponding to the level beyond which further increases in output would lead to a reduction in average cost of no more than 10%. Engineering studies would otherwise often imply an implausibly high value for m.e.s., and one that might lie far above those generated by the other popular measure of m.e.s., viz., the size of the median plant operating in the industry. It was in terms of these definitions that the warranted levels of concentration were seen to be very low compared to actual levels. While this procedure may lead to plausible answers, it actually hides the central analytical issue behind an empirical guess, for it rests upon a judgment as to the size of cost disadvantage a firm may suffer without being driven from the market.

But such questions obviously depend on the intensity of price competition in the market, and this will depend in turn not only on product characteristics (the degree of product differentiation) and on the state of competition policy, etc., but also on the degree of concentration in the industry. But if this is so, it would appear that there is no escaping an explicit analysis of the kind undertaken in chapter 2.

Within the present approach, it is assumed that the firm's average cost (including the contribution of sunk outlays incurred in setting up production) are declining, so that the m.e.s. level—on a strict

definition—is infinite. On the other hand, the toughness of price competition is considered explicitly, and the equilibrium level of concentration is ground out by the interplay between setup costs, market size, and the toughness of price competition.[5]

5. Thus, the warranted degree of concentration does not figure in the present scheme. The use of m.e.s. levels, as estimated from either engineering studies or median plant size, enters only insofar as it is assumed that differences in the cost of establishing a single m.e.s. plant, as defined along traditional lines, may provide a very rough proxy for differences in the level of setup cost across different industries (see chapter 4).