Contestable Markets: An Uprising in the Theory of Industry Structure

William J. Baumol

The American Economic Review, Vol. 72, No. 1. (Mar., 1982), pp. 1-15.

Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28198203%2972%3A1%3C1%3ACMAUIT%3E2.0.CO%3B2-J

The American Economic Review is currently published by American Economic Association.

Contestable Markets: An Uprising in the Theory of Industry Structure

By WILLIAM J. BAUMOL*

The address of the departing president is no place for modesty. Nevertheless, I must resist the temptation to describe the analysis I will report here as anything like a revolution. Perhaps terms such as "rebellion" or "uprising" are rather more apt. But, nevertheless, I shall seek to convince you that the work my colleagues, John Panzar and Robert Willig, and I have carried out and encapsulated in our new book enables us to look at industry structure and behavior in a way that is novel in a number of respects, that it provides a unifying analytical structure to the subject area, and that it offers useful insights for empirical work and for the formulation of policy.

Before getting into the substance of the analysis I admit that this presidential address is most unorthodox in at least one significant respect—that it is not the work of a single author. Here it is not even sufficient to refer to Panzar and Willig, the coauthors of both the substance and the exposition of the book in which the analysis is described in full. For others have made crucial contributions to the formulation of the theory—most notably Elizabeth Bailey, Dietrich Fischer, Herman Quirmbach, and Thijs ten Raa.

But there are many more than these. No uprising by a tiny band of rebels can hope to change an established order, and when the time for rebellion is ripe it seems to break out simultaneously and independently in a variety of disconnected centers each offering its own program for the future. Events here have been no different. I have recently received a proposal for a conference on new developments in the theory of industry structure formulated by my colleague, Joseph Stiglitz, which lists some forty participants, most of them widely known. Among those working on the subject are persons as well known as Caves, Dasgupta, Dixit, Friedlaender, Grossman, Hart, Levin, Ordover, Rosse, Salop, Schmalensee, Sonnenschein, Spence, Varian, von Weizsäcker, and Zeckhauser, among *many* others.[1] It is, of course, tempting to me to take the view that our book is the true gospel of the rebellion and that the doctrines promulgated by others must be combatted as heresy. But that could at best be excused as a manifestation of the excessive zeal one comes to expect on such occasions. In truth, the immediate authors of the work I will report tonight may perhaps be able to justify a claim to have offered some systematization and order to the new doctrines—to have built upon them a more comprehensive statement of the issues and the analysis, and to have made a number of particular contributions. But, in the last analysis, we must look enthusiastically upon our fellow rebels as comrades in arms, each of whom has made a crucial contribution to the common cause.

Turning now to the substance of the theory, let me begin by contrasting our results with those of the standard theory. In offering this contrast, let me emphasize that much of the analysis rests on work that appeared considerably earlier in a variety of forms.

*Presidential address delivered at the ninety-fourth meeting of the American Economic Association, December 29, 1981. I should like to express my deep appreciation to the many colleagues who have contributed to the formulation of the ideas reported here, and to the Economics Program of the Division of Social Sciences of the National Science Foundation, the Division of Information Science and Technology of the National Science Foundation, and the Sloan Foundation for their very generous support of the research that underlies it.

[1]Such a list must inevitably have embarassing omissions—perhaps some of its author's closest friends. I can only say that it is intended just to be suggestive. The fact that it is so far from being complete also indicates how widespread an uprising I am discussing.

William J. Bennett

Number 83 of a series of photographs of past presidents of the Association

We, no less than other writers, owe a heavy debt to predecessors from Bertrand to Bain, from Cournot to Demsetz. Nevertheless, it must surely be acknowledged that the following characterization of the general tenor of the literature as it appeared until fairly recently is essentially accurate.

First, in the received analysis perfect competition serves as the one standard of welfare-maximizing structure and behavior. There is no similar form corresponding to industries in which efficiency calls for a very limited number of firms (though the earlier writings on workable competition did move in that direction in a manner less formal than ours).

Our analysis, in contrast, provides a generalization of the concept of the perfectly competitive market, one which we call a "perfectly contestable market." It is, generally, characterized by optimal behavior and yet applies to the full range of industry structures including even monopoly and oligopoly. In saying this, it must be made clear that perfectly contestable markets do not populate the world of reality any more than perfectly competitive markets do, though there are a number of industries which undoubtedly approximate contestability even if they are far from perfectly competitive. In our analysis, perfect contestability, then, serves not primarily as a description of reality, but as a benchmark for desirable industrial organization which is far more flexible and is applicable far more widely than the one that was available to us before.

Second, in the standard analysis (including that of many of our fellow rebels), the properties of oligopoly models are heavily dependent on the assumed expectations and reaction patterns characterizing the firms that are involved. When there is a change in the assumed nature of these expectations or reactions, the implied behavior of the oligopolistic industry may change drastically.

In our analysis, in the limiting case of perfect contestability, oligopolistic structure and behavior are freed entirely from their previous dependence on the conjectural variations of *incumbents* and, instead, these are generally determined uniquely and, in a manner that is tractable analytically, by the pressures of *potential* competition to which Bain directed our attention so tellingly.

Third, the standard analysis leaves us with the impression that there is a rough continuum, in terms of desirability of industry performance, ranging from unregulated pure monopoly as the pessimal arrangement to perfect competition as the ideal, with relative efficiency in resource allocation increasing monotonically as the number of firms expands.

I will show that, in contrast, in perfectly contestable markets behavior is sharply discontinuous in its welfare attributes. A contestable monopoly offers us some presumption, but no guarantee, of behavior consistent with a second best optimum, subject to the constraint that the firm be viable financially despite the presence of scale economies which render marginal cost pricing financially infeasible. That is, a contestable monopoly has some reason to adopt the Ramsey optimal price-output vector, but it may have other choices open to it. (For the analysis of contestable monopoly, see my article with Elizabeth Bailey and Willig, Panzar and Willig's article, and my book with Panzar and Willig, chs. 7 and 8.)

But once each product obtains a second producer, that is, once we enter the domain of duopoly or oligopoly for each and every good, such choice disappears. The contestable oligopoly which achieves an equilibrium that immunizes it from the incursions of entrants has only one pricing option—it must set its price exactly *equal* to marginal cost and do *all* of the things required for a first best optimum! In short, once we leave the world of pure or partial monopoly, any contestable market must behave ideally in every respect. Optimality is *not* approached gradually as the number of firms supplying a commodity grows. As has long been suggested in Chicago, two firms can be enough to guarantee optimality (see, for example, Eugene Fama and Arthur Laffer).

Thus, the analysis extends enormously the domain in which the invisible hand holds sway. In a perfectly contestable world, it seems to rule almost everywhere. Lest this

seem to be too Panglossian a view of reality, let me offer two observations which make it clear that we emphatically do not believe that all need be for the best in this best of all possible worlds.

First, let me recall the observation that real markets are rarely, if ever, perfectly contestable. Contestability is merely a broader ideal, a benchmark of wider applicability than is perfect competition. To say that contestable oligopolies behave ideally and that contestable monopolies have some incentives for doing so is not to imply that this is even nearly true of all oligopolies or of unregulated monopolies in reality.

Second, while the theory extends the domain of the invisible hand in some directions, it unexpectedly restricts it in others. This brings me to the penultimate contrast I wish to offer here between the earlier views and those that emerge from our analysis.

The older theoretical analysis seems to have considered the invisible hand to be a rather weak intratemporal allocator of resources, as we have seen. The mere presence of unregulated monopoly or oligopoly was taken to be sufficient per se to imply that resources are likely to be misallocated *within* a given time period. But *where the market structure is such as to yield a satisfactory allocation of resources within the period*, it may have seemed that it can, at least in theory, do a good job of intertemporal resource allocation. In the absence of any externalities, persistent and asymmetric information gaps, and of interference with the workings of capital markets, the amounts that will be invested for the future may appear to be consistent with Pareto optimality and efficiency in the supply of outputs to current and future generations.

However, our analysis shows that where there are economies of scale in the production of durable capital, intertemporal contestable monopoly, which may perform relatively well in the single period, cannot be depended upon to perform ideally as time passes. In particular, we will see that the least costly producer is in the long run vulnerable to entry or replacement by rivals whose appearance is inefficient because it wastes valuable social resources.

There is one last contrast between the newer analyses and the older theory which I am most anxious to emphasize. In the older theory, the nature of the industry structure was *not* normally explained by the analysis. It was, in effect, taken to be given exogenously, with the fates determining, apparently capriciously, that one industry will be organized as an oligopoly, another as a monopoly and a third as a set of monopolistic competitors. Assuming that this destiny had somehow been revealed, the older analyses proceeded to investigate the consequences of the exogenously given industry structure for pricing, outputs, and other decisions.[2]

The new analyses are radically different in this respect. In our analysis, among others, an industry's structure is determined explicitly, endogenously, and simultaneously with the pricing, output, advertising, and other decisions of the firms of which it is constituted. This, perhaps, is one of the prime contributions of the new theoretical analyses.

I. Characteristics of Contestable Markets

Perhaps a misplaced instinct for melodrama has led me to say so much about contestable markets without even hinting what makes a market contestable. But I can postpone the definition no longer. A contestable market is one into which entry is absolutely free, *and exit is absolutely costless.* We use "freedom of entry" in Stigler's sense, not to mean that it is costless or easy, but that the entrant suffers no disadvantage in terms of production technique or perceived product quality relative to the incumbent,

[2] Of course, any analysis which considered the role of entry, whether it dealt with perfect competition or monopolistic competition, must implicitly have considered the determination of industry structure by the market. But in writings before the 1970's, such analyses usually did not consider how this process determined whether the industry would or would not turn out to be, for example, an oligopoly. The entry conditions were studied only to show how the *assumed* market structure could constitute an equilibrium state. Many recent writings have gone more explicitly into the determination of industry structure, though their approaches generally differ from ours.

and that potential entrants find it appropriate to evaluate the profitability of entry in terms of the incumbent firms' pre-entry prices. In short, it is a requirement of contestability that there be no cost discrimination against entrants. Absolute freedom of exit, to us, is one way to guarantee freedom of entry. By this we mean that any firm can leave without impediment, and in the process of departure can recoup any costs incurred in the entry process. If all capital is salable or reusable without loss other than that corresponding to normal user cost and depreciation, then any risk of entry is eliminated.

Thus, contestable markets may share at most one attribute with perfect competition. Their firms need not be small or numerous or independent in their decision making or produce homogeneous products. In short, a perfectly competitive market is necessarily perfectly contestable, but not *vice versa*.

The crucial feature of a contestable market is its vulnerability to hit-and-run entry. Even a very transient profit opportunity need not be neglected by a potential entrant, for he can go in, and, before prices change, collect his gains and then depart without cost, should the climate grow hostile.

Shortage of time forces me to deal rather briefly with two of the most important properties of contestable markets—their welfare attributes and the way in which they determine industry structure. I deal with these briefly because an intuitive view of the logic of these parts of the analysis is not difficult to provide. Then I can devote a bit more time to some details of the oligopoly and the intertemporal models.

A. *Perfect Contestability and Welfare*

The welfare properties of contestable markets follow almost directly from their definition and their vulnerability to hit-and-run incursions. Let me list some of these properties and discuss them succinctly.

First, a contestable market never offers more than a normal rate of profit—its economic profits must be zero or negative, even if it is oligopolistic or monopolistic. The reason is simple. Any positive profit means that a transient entrant can set up business, replicate a profit-making incumbent's output at the same cost as his, undercut the incumbent's prices slightly and still earn a profit. That is, continuity and the opportunity for costless entry and exit guarantee that an entrant who is content to accept a slightly lower economic profit can do so by selecting prices a bit lower than the incumbent's.

In sum, in a perfectly contestable market any economic profit earned by an incumbent automatically constitutes an earnings opportunity for an entrant who will hit and, if necessary, run (counting his temporary but supernormal profits on the way to the bank). Consequently, in contestable markets, zero profits must characterize any equilibrium, even under monopoly and oligopoly.

The second welfare characteristic of a contestable market follows from the same argument as the first. This second attribute of any contestable market is the absence of any sort of inefficiency in production in industry equilibrium. This is true alike of inefficiency of allocation of inputs, *X*-inefficiency, inefficient operation of the firm, or inefficient organization of the industry. For any unnecessary cost, like any abnormal profit, constitutes an invitation to entry. Of course, in the short run, as is true under perfect competition, both profits and waste may be present. But in the long run, these simply cannot withstand the threat brandished by potential entrants who have nothing to lose by grabbing at any opportunity for profit, however transient it may be.

A third welfare attribute of any long-run equilibrium in a contestable market is that no product can be sold at a price, p, that is less than its marginal cost. For if some firm sells y units of output at such a price and makes a profit in the process, then it is possible for an entrant to offer to sell a slightly smaller quantity, $y-\varepsilon$, at a price a shade lower than the incumbent's, and still make a profit. That is, if the price p is less than MC, then the sale of $y-\varepsilon$ units at price p must yield a total profit $\pi+\Delta\pi$ which is greater than the profit, π, that can be earned by selling only y units of output at that price. Therefore, there must exist a price just slightly lower than p which enables the entrant to undercut the incumbent and yet to

earn at least as much as the incumbent, by eliminating the unprofitable marginal unit.

This last attribute of contestable equilibria —the fact that price must always at least equal marginal cost—is important for the economics of antitrust and regulation. For it means that in a perfectly contestable market, no cross subsidy is possible, that is, no predatory pricing can be used as a weapon of unfair competition. But we will see it also has implications which are more profound theoretically and which are more germane to our purposes. For it constitutes half of the argument which shows that when there are two or more suppliers of any product, its price must, in equilibrium, be exactly equal to marginal cost, and so resource allocation must satisfy all the requirements of first best optimality.

Indeed, the argument here is similar to the one which has just been described. But there is a complication which is what introduces the two-firm requirement into this proposition. $p < MC$ constitutes an opportunity for profit to an entrant who drops the unprofitable marginal unit of output, as we have just seen. It would seem, symmetrically, that $p > MC$ also automatically constitutes an opportunity for profitable entry. Instead of selling the y-unit output of a profitable incumbent, the entrant can now offer to sell the slightly larger output, $y + \varepsilon$, using the profits generated by the marginal unit at a price greater than marginal cost to permit a reduction in price below the incumbent's. But on this side of the incumbent's output, there is a catch in the argument. Suppose the incumbent is a monopolist. Then output and price are constrained by the elasticity of demand. An attempt by an entrant to sell $y + \varepsilon$ rather than y may conceivably cause a sharp reduction in price which eliminates the apparent profits of entry. In the extreme case where demand is perfectly inelastic, there will be no positive price at which the market will absorb the quantity $y + \varepsilon$. This means that the profit opportunity represented by $p > MC$ can crumble into dust as soon as anyone seeks to take advantage of it.

But all this changes when the market contains two or more sellers. Now $p > MC$ does always constitute a real opportunity for profitable entry. The entrant who wishes to sell a bit more than some one of the profitable incumbents, call him incumbent A, need not press against the industry's total demand curve for the product. Rather, he can undercut A, steal away all of his customers, at least temporarily, and, in addition, steal away ε units of demand from any other incumbent, B. Thus, if A and B together sell $y_a + y_b > y_a$, then an entrant can lure away $y_a + \varepsilon > y_a$ customers, for ε sufficiently small, and earn on this the incremental profit $\varepsilon(p - MC) > 0$. This means that the entrant who sells $y_a + \varepsilon$ can afford to undercut the prevailing prices somewhat and still make more profit than an incumbent who sells y_a at price p.

In sum, where a product is sold by two or more firms, any $p > MC$ constitutes an irresistible entry opportunity for hit-and-run entry in a perfectly contestable market, for it promises the entrant supernormal profits even if they accrue for a very short period of time.

Consequently, when a perfectly contestable market contains two or more sellers, neither $p < MC$ nor $p > MC$ is compatible with equilibrium. Thus we have our third and perhaps most crucial welfare attribute of such perfectly contestable markets—their prices, in equilibrium, must be equal to marginal costs, as is required for Pareto optimality of the "first best" variety. This, along with the conclusion that such markets permit no economic profits and no inefficiency in long-run equilibrium, constitutes their critical properties from the viewpoint of economic welfare. Certainly, since they do enjoy those three properties, the optimality of perfectly contestable equilibria (with the reservations already expressed about the case of pure monopoly) fully justifies our conclusion that perfect contestability constitutes a proper generalization of the concept of perfect competition so far as welfare implications are concerned.

B. *On the Determination of Industry Structure*

I shall be briefer and even less rigorous in describing how industry structure is determined endogenously by contestability

analysis. Though this area encompasses one of its most crucial accomplishments, there is no way I can do justice to the details of the analysis in an oral presentation and within my allotted span of time. However, an intuitive view of the matter is not difficult.

The key to the analysis lies in the second welfare property of contestable equilibria—their incompatibility with inefficiency of any sort. In particular, they are incompatible with inefficiency in the *organization* of an industry. That is, suppose we consider whether a particular output quantity of an industry will be produced by two firms or by a thousand. Suppose it turns out that the two-firm arrangement can produce the given output at a cost 20 percent lower than it can be done by the 1,000 firms. Then one implication of our analysis is that the industry cannot be in long-run equilibrium if it encompasses 1,000 producers. Thus we already have some hint about the equilibrium industry structure of a contestable market.

We can go further with this example. Suppose that, with the given output vector for the industry, it turns out that *no* number of firms other than two can produce at as low a total cost as is possible under a two-firm arrangement. That is, suppose two firms can produce the output vector at a total cost lower than it can be done by one firm or three firms or sixty or six thousand. Then we say that for the given output vector the industry is a *natural duopoly*.

This now tells us how the industry's structure can be determined. We proceed, conceptually, in two steps. First we determine what structure happens to be most efficient for the production of a given output vector by a given industry. Next, we investigate when market pressures will lead the industry toward such an efficient structure in equilibrium.

Now, the first step, though it has many intriguing analytic attributes, is essentially a pure matter of computation. Given the cost function for a typical firm, it is ultimately a matter of calculation to determine how many firms will produce a given output most efficiently. For example, if economies of scale hold throughout the relevant range and there are sufficient complementarities in the production of the different commodities supplied by the firm, then it is an old and well-known conclusion that single firm production will be most economical—that we are dealing with a natural monopoly.

Similarly, in the single product case suppose the average cost curve is U shaped and attains its minimum point at an output of 10,000 units per year. Then it is obvious that if the industry happens to sell 50,000 units per year, this output can be produced most cheaply if it is composed of exactly five firms, each producing 10,000 units at its point of minimum average cost.

Things become far more complex and more interesting when the firm and the industry produce a multiplicity of commodities, as they always do in reality. But the logic is always the same. When the industry output vector is small compared to the output vectors the firm can produce at relatively low cost, then the efficient industry structure will be characterized by very few firms. The opposite will be true when the industry's output vector is relatively far from the origin. In the multiproduct case, since average cost cannot be defined, two complications beset the characterization of the output vectors which the firm can produce relatively efficiently. First, since here average cost cannot be defined, we cannot simply look for the point of minimum average costs. But we overcome this problem by dealing with output bundles having fixed proportions among commodity quantities—by moving along a ray in output space. Along any such ray the behavior of average cost *is* definable, and the point of minimum ray average cost (RAC) is our criterion of relatively efficient scale for the firm. Thus, in Figure 1 we have a ray average cost curve for the production of boots and shoes when they are produced in the proportion given by ray $0R$. We see that for such bundles y^m is the point of minimum RAC. A second problem affecting the determination of the output vectors the firm can produce efficiently is the choice of output proportions—the location of the ray along which the firm will operate. This depends on the degree of complementarity in production of the goods, and it also lends itself to formal analysis.

We note also that the most efficient number of firms will vary with the location of the

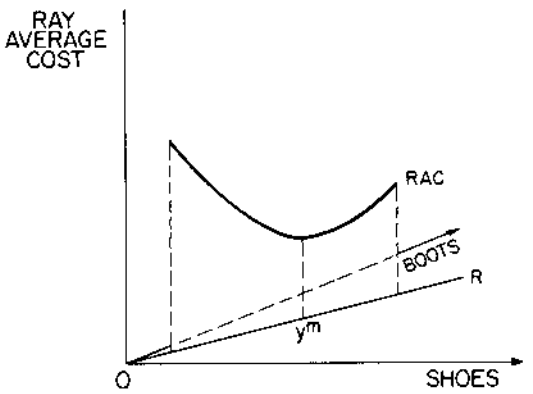

FIGURE 1

industry's output vector. The industry may be a natural monopoly with one output vector, a natural duopoly with another, and efficiency may require seventy-three firms when some third output vector is provided by the industry.

This, then, completes the first of the two basic steps in the endogenous determination of industry structure. Here we have examined what industry structure is least costly for each given output vector of a given industry, and have found how the result depends on the magnitudes of the elements of that output vector and the shape of the cost function of the typical firm. So far the discussion may perhaps be considered normative rather than behavioral. It tells us what structure is most efficient under the circumstances, not which industry structure will emerge under the pressures of the market mechanism.

The transition toward the second, behavioral, stage of the analysis is provided by the observation that the optimal structure of an industry depends on its output vector, while that output vector in turn depends on the prices charged by its firms. But, since pricing depends on industry structure, we are brought full circle to the conclusion that pricing behavior and industry structure must, ultimately, be determined simultaneously and endogenously.

We are in no position to go much further than this for a market whose properties are unspecified. But, for a perfectly contestable market, we can go much further. Indeed, the properties of perfect contestability cut through every difficulty and tell us the equilibrium prices, outputs, and industry structure, all at once.

Where more than one firm supplies a product, we have already characterized these prices precisely. For we have concluded that each equilibrium price will equal the associated marginal cost. Then, given the industry's cost and demand relationships, this yields the industry's output quantities simultaneously with its prices, in the usual manner. Here there is absolutely nothing new in the analysis.

But what is new is the format of the analysis of the determination of industry structure. As I have already pointed out, structure is determined by the efficiency requirement of equilibrium in any contestable market. Since no such equilibrium is compatible with failure to minimize industry costs, it follows that the market forces under perfect contestability will bring us results consistent with those of our normative analysis. Whatever industry structures minimize total costs for the equilibrium output vector must turn out to be the only structures consistent with industry equilibrium in the long run.

Thus, for contestable markets, but for contestable markets *only*, the second stage of the analysis of industry structure turns out to be a sham. Whatever industry structure was shown by the first, normative, portion of the analysis to be least costly must also emerge as the industry structure selected by market behavior. No additional calculations are required by the behavioral analysis. It will all have been done in the normative cost-minimization analysis and the behavioral analysis is pure bonus.

Thus, as I promised, I have indicated how contestability theory departs from the older theory which implicitly took industry structure to be determined exogenously in a manner totally unspecified and, instead, along with other recent writings, embraces the determination of industry structure as an integral part of the theory to be dealt with simultaneously with the determination of prices and outputs.

At this point I can only conjecture about the determination of industry structure once we leave the limiting case of perfect contestability. But my guess is that there are no

sharp discontinuities here, and that while the industry structures which emerge in reality are not always those which minimize costs, they will constitute reasonable approximations to the efficient structures. If this is not so it is difficult to account for the similarities in the patterns of industry structure that one observes in different countries. Why else do we not see agriculture organized as an oligopoly in any free market economy, or automobiles produced by 10,000 firms? Market pressures must surely make any very inefficient market structure vulnerable to entry, to displacement of incumbents by foreign competition, or to undermining in other ways. If that is so, the market structure that is called for by contestability theory may not prove to be too bad an approximation to what we encounter in reality.

II. On Oligopoly Equilibrium

I should like now to examine oligopoly equilibrium somewhat more extensively. We have seen that, except where a multiproduct oligopoly firm happens to sell some of its products in markets in which it has no competitors, an important partial monopoly case which I will ignore in what follows, all prices must equal the corresponding marginal costs in long-run equilibrium. But in an oligopoly market, this is a troublesome concept. Unless the industry output vector happens to fall at a point where the cost function is characterized by locally constant returns to scale, we know that zero profits are incompatible with marginal cost pricing. Particularly if there are scale economies at that point, so that marginal cost pricing precludes financial viability, we can hardly expect such a solution to constitute an equilibrium. Besides, we have seen that long-run equilibrium requires profit to be precisely zero. We would thus appear to have run into a major snag by concluding that perfect contestability always leads to marginal cost pricing under oligopoly.

This is particularly so if the (ray) average curve is U shaped, with its minimum occurring at a single point, y^m. For in this case that minimum point is the only output of the firm consistent with constant returns to scale and with zero profits under marginal cost pricing. Thus, dealing with the single product case to make the point, it would appear, say, that if the AC-minimizing output is 1,000, in a contestable market, equilibrium is possible if quantity demanded from the industry happens to be exactly 2,000 units (so two firms can produce 1,000 units each) or exactly 3,000 units or exactly 4,000 units, etc. But suppose the demand curve happens to intersect the industry AC curve, say, at 4,030 units. That is, then, the only industry output satisfying the equilibrium requirement that price equals zero profit. But then, at least one of the four or five firms in the industry must produce either more or less than 1,000 units of output, and so the slope of its AC curve will not be zero at that point, precluding either MC pricing or zero profits and, consequently, violating one or the other of the requirements of equilibrium in a perfectly contestable market.

It would appear that equilibrium will be impossible in this perfectly contestable market unless by a great piece of luck the industry demand curve happens to intersect its AC curve at 2,000 or 3,000 units or some other integer multiple of 1,000 units of output.

There are a variety of ways in which one can grapple with this difficulty. In his dissertation at New York University, Thijs ten Raa has explored the issue with some care and has shown that the presence of entry costs of sufficient magnitude, that is, irreversible costs which must be borne by an entrant but not by an incumbent, can eliminate the existence problem. The minimum size of the entry cost required to permit an equilibrium will depend on the size of the deviation from zero profits under marginal cost pricing and ten Raa has given us rules for its determination. He has shown also that the existence problem, as measured by the required minimum size of entry cost, decreases rapidly as the equilibrium number of firms of the industry increases, typically attaining negligible proportions as that number reaches, say, ten enterprises. For, as is well known, when the firm's average cost curve is U shaped the industry's average cost curve will approach a horizontal line as the

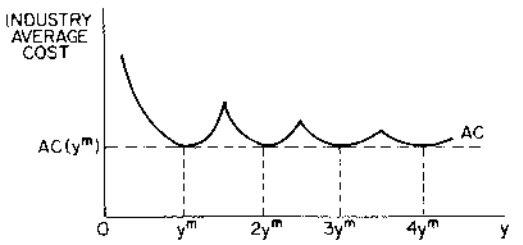

FIGURE 2

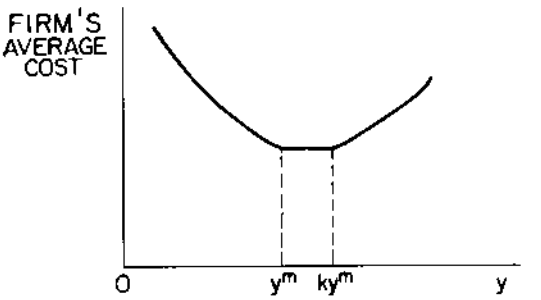

FIGURE 3

size of industry output increases. This is shown in Figure 2 which is a standard diagram giving the firm's and the industry's AC curves when the former is U shaped. As a result, the deviations between average cost and marginal cost will decline as industry output increases and so the minimum size of the entry cost required to preserve equilibrium declines correspondingly.

However, here I want to describe another approach offered in our book to the problem of existence which I have just described—the difficulty of satisfying simultaneously the zero-profit requirement and the requirement of marginal cost pricing. This second avenue relies on the apparently unanimous conclusion of empirical investigators of the cost function of the firm, that AC curves are not, in fact, characterized by a unique minimum point as they would be if they had a smooth U shape. Rather, these investigators tell us, the AC curve of reality has a flat bottom—an interval along which it is horizontal. That is, average costs do tend to fall at first with size of output, then they reach a minimum and continue at that level for some range of outputs, after which they may begin to rise once more. An AC curve of this variety is shown in Figure 3. Obviously, such a flat segment of the AC curves *does* help matters because there is now a *range* of outputs over which MC pricing yields zero profits. Moreover, the longer the flat-bottomed segment the better matters are for existence of equilibrium. Indeed, it is easy to show that if the left-hand end of the flat segment occurs at output y^m and the right-hand end occurs at ky_m, then if k is greater than or equal to 2 the existence problem disappears altogether, because the industry's AC curves will be horizontal for any output greater than y_m. That is, in any contestable market in which two or more firms operate the industry AC curve will be horizontal and MC pricing will always yield zero profits. To confirm that this is so, note that if, for example, the flat segment for the firm extends from $y = 1{,}000$ to $y = 2{,}000$, then any industry output of, say, $9{,}000 + \Delta y$ where $0 \leqslant \Delta y \leqslant 9{,}000$ can be produced by nine firms, each of them turning out more than 1,000 but less than 2,000 units. Hence, each of them will operate along the horizontal portion of its AC curve, as equilibrium requires.

Thus, if the horizontal interval (y^m, ky_m) happens to satisfy $k \geqslant 2$, there is no longer any problem for existence of equilibrium in a contestable market with two or more firms. But fate may not always be so kind. What if that horizontal interval is quite short, that is, k is quite close to unity? Such a case is shown in our diagram where for illustration I have taken $k = 4/3$.

I should like to take advantage of your patience by dealing here not with the simplest case—that of the single product industry—but with the multiproduct problem. I do this partly to offer you some feeling of the way in which the multiproduct analysis, which is one of the hallmarks of our study, works out in practice.

Because, as we have seen, there is no way one can measure average cost for all output combinations in the multiproduct case, I will deal exclusively with the total cost function. Figure 4 shows such a total cost function for the single firm, which is taken to manufacture two products, boots and shoes.

Let us pause briefly to examine its shape. Along any ray such as OR, which keeps

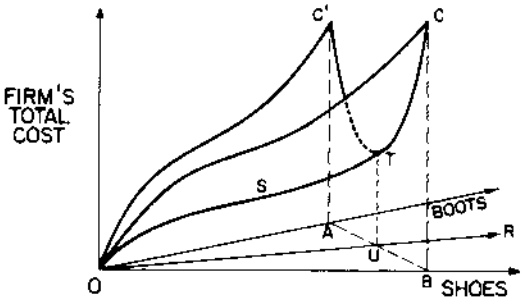

FIGURE 4

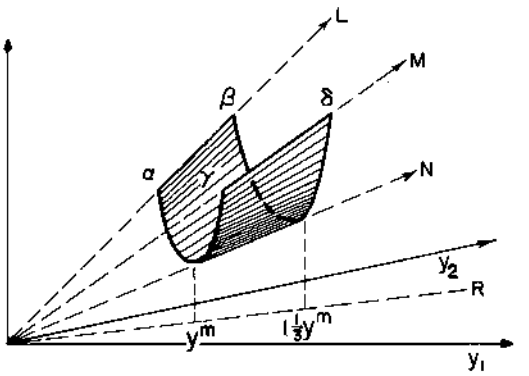

FIGURE 5

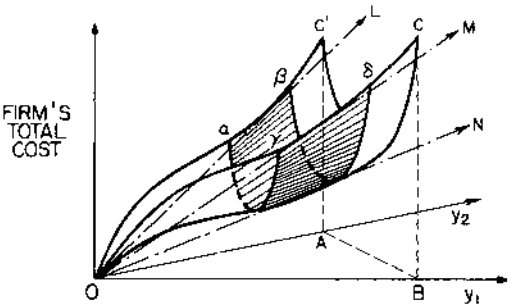

FIGURE 6

output proportions constant, we have an ordinary total cost curve, *OST*. With one exception, which I will note soon, I have drawn it to have the usual sort of shape, with marginal costs falling near the origin and rising at points much further from the origin. On the other hand, the trans ray cut above *AB* yields a cross section *C'TC* which is more or less U shaped. This means that it is relatively cheaper to produce boots and shoes together (point *U*) than to produce them in isolation (point *A* or point *B*). That is, this convex trans ray shape is enough to offer us the complementarity which leads firms and industries to turn out a multiplicity of products rather than specializing in the production of a single good.

Now what, in such a case, corresponds to the flat bottom of an *AC* curve in a single product case? The answer is that the cost function in the neighborhood of the corresponding output must be linearly homogeneous. In Figure 5 such a region, $\alpha\beta\gamma\delta$, is depicted. It is linearly homogeneous because it is generated by a set of rays such as *L*, *M*, and *N*. For simplicity in the discussion that follows, I have given this region a very regular shape—it is, approximately, a rectangle which has been moved into three-dimensional space and given a U-shaped cross section.

Now Figure 6 combines the two preceding diagrams and we see that they have been drawn to mesh together, so that the linearly homogeneous region constitutes a portion of the firm's total cost surface. We see then that the firm's total cost does have a region in which constant returns to scale occur, and which corresponds to the flat-bottomed segment of the *AC* curve.

Moreover, as before, I have deliberately kept this segment quite narrow. Indeed, I have repeated the previous proportions, letting the segment extend from a distance y^m from the origin to the distance $1\frac{1}{3}y^m$ along any ray on the floor of the diagram.

Let us now see what happens in these circumstances when we turn to the total cost surface for the *industry*. This is depicted in Figure 7 which shows a relationship that may at first seem surprising. In Figure 7 I depict only the linearly homogeneous portions of the industry's cost surface. There we see that while for the firm linear homogeneity prevailed only in the interval from y^m to $1\frac{1}{3}y^m$, in the case of industry output linear homogeneity also holds in that same interval but, in addition, it holds for the interval $2y^m$ to $2\frac{2}{3}y^m$ and in the region extending from $3y^m$ to infinity. That is, everywhere beyond $3y^m$ the industry's total cost function is linearly homogeneous. In this case, then, we have three regions of local linear homogene-

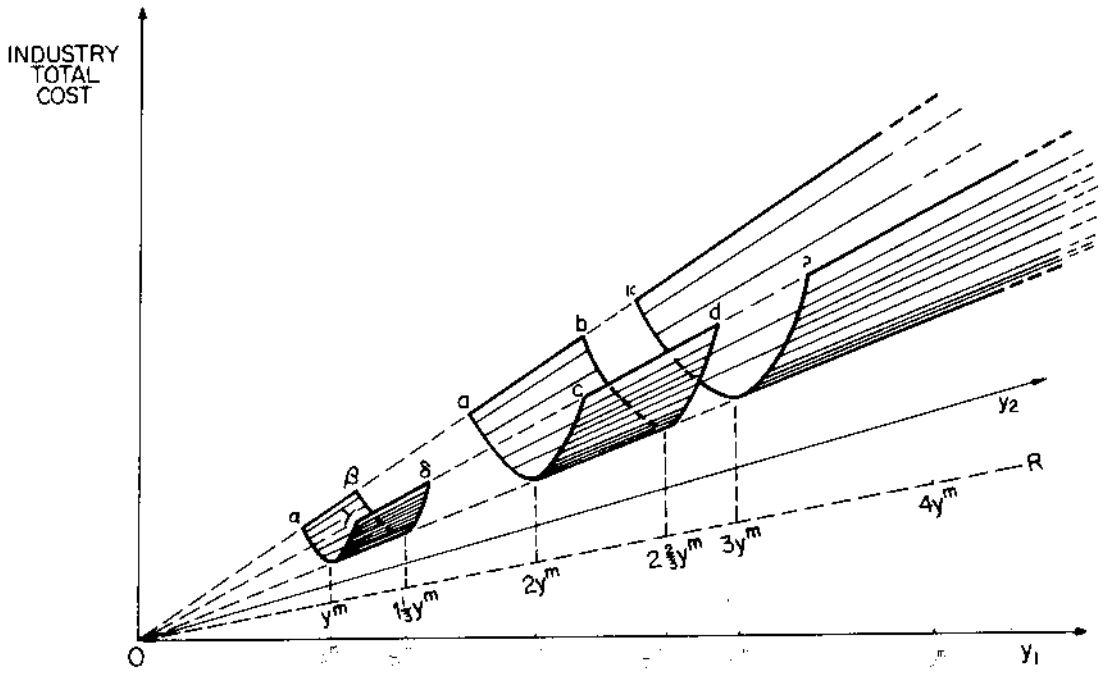

FIGURE 7

ity in the industry's cost function, $\alpha\beta\gamma\delta$, which is identical with that of the individual firm, the larger region *abcd*, and the infinite region *aleph beth*....

Before showing why this is so we must pause to note the implications of the exercise. For it means that even a relatively small region of flatness in the *AC* curve of the individual firm, that is, of linear homogeneity in its total cost function, eliminates the bulk of the existence problem for oligopoly equilibrium in a contestable market. The problem does not arise for outputs nearer to the origin than y_m because such outputs are supplied most efficiently by a monopoly which is not required to price at marginal cost in a contestable market equilibrium. The problem also does not arise for any industry output greater than $3y^m$ in this case, because everywhere beyond that marginal cost pricing yields zero profits. There are two relatively narrow regions in which no equilibrium is, indeed, possible, but here we may conjecture that the vicissitudes of disequilibrium will cause shifts in the demand relationships as changing prices and changing consumption patterns affect tastes, and so the industry will ultimately happen upon an equilibrium position and remain there until exogenous disturbances move it away. Thus we end up with an oligopoly equilibrium whose prices, profits, and other attributes are determined without benefit of the conjectural variation, reaction functions, and the other paraphernalia of standard oligopoly analysis.

To complete this discussion of oligopoly equilibrium in a contestable market, it only remains for me to explain why the regions of linear homogeneity in the industry's cost function are as depicted in Figure 7. The answer is straightforward. Let $C(y)$ be the firm's total cost function for which we have assumed for expository simplicity that in the interval from y^m to $1\frac{1}{3}y^m$ along each and every ray, total cost grows exactly proportionately with output. Then two firms can produce $2y^m$ at the same unit cost, and three firms can produce $3y^m$ at that same unit cost for the given output bundle, etc. But by exactly the same argument, the two firms together, each producing no more than $1\frac{1}{3}y^m$,

can turn out anything up to $2\frac{2}{3}y^m$ without affecting unit costs, and three firms can produce as much as $3\frac{3}{3}y^m$, that is, as much as $4y^m$. In sum, the intervals of linear homogeneity for the industry are the following:

Interval 1: from y^m to $1\frac{1}{3}y^m$
Interval 2: from $2y^m$ to $2\frac{2}{3}y^m$
Interval 3: from $3y^m$ to $4y^m$
Interval 4: from $4y^m$ to $5\frac{1}{3}y^m$
Interval 5: from $5y^m$ to $6\frac{2}{3}y^m$
. . . . . .

That is, each interval begins at an integer multiple of y^m and extends $1/3\ y^m$ further than its predecessor. Thus, beyond $3y^m$ successive intervals begin to touch or overlap and that is why linear homogeneity extends everywhere beyond $3y^m$ as I claimed.[3]

There is one complication in the multiproduct case which I have deliberately slid over, feeling the discussion was already complicated enough. The preceding argument assumes implicitly that the firms producing the industry output all employ the same output proportions as those in the industry output vector. For otherwise, it is not legitimate to move outward along a single ray as the number of firms is increased. But suppose increased industry output were to permit savings through increased specialization. Might there not be constant returns with fixed output proportions and yet economies of scale for the industry overall? This problem is avoided by our complementarity assumption used to account for the industry's multiproduct operation—our U-shaped trans-ray cross section. This, in effect, rules out such savings from specialization in the regions where linear homogeneity also rules out savings from increased scale.

This, then, completes my discussion of oligopoly equilibrium in perfectly contestable markets, which we have seen, yields a determinate set of prices and outputs that is not dependent upon assumptions about the nature of incumbent firm's expectations relating to entrants' behavior and offers us a concrete and favorable conclusion on the welfare implications of contestable oligopoly.

[3]The reader can readily generalize this result. If the flat-bottomed segment for the firm extends from y^m to $y^m(1+1/w)$, where w is an integer, then there will be w regions of linear homogeneity in the industry cost function and it will be linearly homogeneous for any output $y \geqslant wy^m$.

III. Intertemporal Vulnerability to Inefficient Entry

Having so far directed attention to areas in which the invisible hand manifests unexpected strength, I should like to end my story by dealing with an issue in relation to which it is weaker than some of us might have expected. As I indicated before, this is the issue of intertemporal production involving durable capital goods.

The analysis is far more general than the following story suggests, but even the case I describe is sufficiently general to make the point. We deal with an industry in which a product is offered by a single firm that provides it period after period. The equilibrium quantity of the commodity that is demanded grows steadily with the passage of time in a manner that is foreseen without uncertainty. Because of economies of scale in the production of capacity the firm deliberately builds some excess capacity to take care of anticipated growth in sales volume. But there is some point, let us say, $z=45$ years in the future, such that it would be uneconomic to take further growth in sales volume into account in the initial choice of capacity. This is so because the opportunity (interest) cost of the capacity that remains idle for 45 or more years exceeds the savings made possible by the economies of scale of construction. Thus, after 45 years it will pay the firm to undertake a second construction project to build the added capacity needed to produce the goods demanded of it.

Suppose that in every particular period our producer is a natural monopolist, that is, he produces the industry's supply of its one commodity at a cost lower than it can be done by any two or more enterprises. Then considering that same product in different periods to be formally equivalent to different goods we may take our supplier to be an intertemporal natural monopolist in a multiproduct industry. That is, no combination of

two or more firms can produce the industry's intertemporal output vector as cheaply as he. I will prove now under a set of remarkably unrestrictive assumptions that despite its cost advantages, there exists no intertemporal price vector consistent with equilibrium for this firm. That is, whatever his price vector, his market will at some time be vulnerable to partial or complete takeover by an entrant who has neither superior skills nor technological superiority and whose entrance increases the quantities of resources used up in production. In other words, here the invisible hand proves incapable of protecting the most efficient producing arrangement and leaves the incumbent producer vulnerable to displacement by an aggressive entrant. I leave to your imaginations what, if anything, this says about the successive displacements on the world market of the Dutch by the English, the English by the Germans and the Americans, and the Americans, perhaps, by the Japanese.

The proof of our proposition on the intertemporal vulnerability of incumbents to entry that is premature from the viewpoint of cost minimization does require just a little bit of algebra. To keep our analysis simple, I will divide time into two periods, each lasting $z = 45$ years so that capacity in the first period is, optimally, just sufficient to satisfy all demand, but in the second, it requires the construction of added capacity to meet demand growth because, by assumption, anticipatory construction to meet growth more than z years in the future simply is too costly. Also for simplicity, I will assume that there are no costs other than cost of construction. Of course, neither this nor the use of only two periods really affects the argument in any way. My only three substantive assumptions are that demand is growing with time, that there are economies of scale, that is, declining average costs in construction, and that there exists some length of time, z, so great that it does not pay in the initial construction to build capacity sufficient for the growth in quantity demanded that will occur beyond that date.

The argument, like the notation, is now straightforward. Let y_t be output in period t, p_t be price in period t, and $K(y)$ be the cost of construction of capacity sufficient to produce (a maximum of) y units per period. Here, both p_t and $K(y)$ are expressed in discounted present value.[4]

Then, by assumption, our firm will construct at the beginning of the first period capacity just sufficient to produce output y_1 at cost $K(y_1)$ and at the beginning of the second period it will produce the rest of the capacity it needs, $y_2 - y_1 > 0$, at the cost $K(y_2 - y_1)$.

The first requirement for the prices in question to be consistent with equilibrium is that they permit the incumbent to cover his costs, that is, that

$$(1) \quad p_1 y_1 + p_2 y_2 \geq K(y_1) + K(y_2 - y_1).$$

Second, for these prices to constitute an equilibrium they must protect the incumbent against any and all possible incursions by entrants. That is, suppose an entrant were to consider the possibility of constructing capacity y_1 and not expanding in the future, and, by undercutting the incumbent, selling the same output, y_1, in each period. Entry on these terms will in fact be profitable unless the prices are such that the sale of y_1 in each period does not bring in revenues sufficient to cover the cost, $K(y_1)$, of the entrant's once-and-for-all construction. That is, entry will be profitable unless

$$(2) \quad p_1 y_1 + p_2 y_1 \leq K(y_1).$$

Thus, the prices in question cannot constitute an equilibrium unless (2) as well as (1) are satisfied.

Now, subtracting (2) from (1) we obtain immediately

$$p_2(y_2 - y_1) \geq K(y_2 - y_1)$$

or

$$(3) \quad p_2 \geq K(y_2 - y_1)/(y_2 - y_1),$$

[4] That is, if p_1^*, p_2^*, represent the undiscounted prices, $p_1 = p_1^*$, $p_2 = p_2^*/(1 + r)$, where r is the rate of interest, etc.

but, by the assumption that average construction cost is declining, since $y_1 > 0$,

$$(4) \quad K(y_2 - y_1)/(y_2 - y_1) > K(y_2)/y_2.$$

Substituting this into (3) we have at once

$$p_2 > K(y_2)/y_2$$

or

$$(5) \quad p_2 y_2 > K(y_2).$$

Inequality (5) is our result. For it proves that any prices which satisfy equilibrium requirements (1) and (2) must permit a second-period entrant using the same techniques to build capacity y_2 from the ground up, at cost $K(y_2)$, to price slightly below anything the incumbent can charge and yet recover his costs; and that in doing so, the entrant can earn a profit.

Thus, our intertemporal natural monopolist cannot quote, *at time zero*, any prices capable of preventing the takeover of some or all of his market. Moreover, this is so despite the waste, in the form of replication of the incumbent's plant, that this entails. That, then, is the end of the formal argument, the proof that here the invisible hand manifests weakness that is, perhaps, unexpected.

You will all undoubtedly recognize that the story as told here in its barest outlines omits all sorts of nuances, such as entrants' fear of responsive pricing, the role of bankruptcy, depreciation of capital, and the like. This is not the place to go into these matters for it is neither possible nor appropriate here for me to go beyond illustration of the logic of the new analysis.

IV. Concluding Comments

Before closing let me add a word on policy implications, whose details must also be left to another place. In spirit, the policy conclusions are consistent with many of those economists have long been espousing. At least in the intratemporal analysis, the heroes are the (unidentified) potential entrants who exercise discipline over the incumbent, and who do so most effectively when entry is free. In the limit, when entry and exit are completely free, efficient incumbent monopolists and oligopolists may in fact be able to prevent entry. But they can do so only by behaving virtuously, that is, by offering to consumers the benefits which competition would otherwise bring. For every deviation from good behavior instantly makes them vulnerable to hit-and-run entry.

This immediately offers what may be a new insight on antitrust policy. It tells us that a history of absence of entry in an industry and a high concentration index may be signs of virtue, not of vice. This will be true when entry costs in our sense are negligible. And, then, efforts to change market structure must be regarded as mischievous and antisocial in their effects.

A second and more obvious conclusion is the questionable desirability of artificial impediments to entry, such as regulators were long inclined to impose. The new analysis merely reinforces the view that any proposed regulatory barrier to entry must start off with a heavy presumption against its adoption. Perhaps a bit newer is the emphasis on the importance of freedom of exit which is as crucial a requirement of contestability as is freedom of entry. Thus we must reject as perverse the propensity of regulators to resist the closing down of unprofitable lines of activity. This has even gone so far as a Congressional proposal (apparently supported by Ralph Nader) to require any plant with yearly sales exceeding $250,000 to provide fifty-two weeks of severance pay and to pay three years of taxes, before it will be permitted to close, and that only after giving two years notice!

There is much more to the policy implications of the new theory, but I will stop here, also leaving its results relating to empirical research for discussion elsewhere.

Let me only say in closing that I hope I have adequately justified my characterization of the new theory as a rebellion or an uprising. I believe it offers a host of new analytical methods, new tasks for empirical research, and new results. It permits reexamination of the domain of the invisible hand, yields contributions to the theory of

oligopoly, provides a standard for policy that is far broader and more widely applicable than that of perfect competition, and leads to a theory that analyzes the determination of industry structure endogenously and simultaneously with the analysis of the other variables more traditionally treated in the theory of the firm and the industry. It aspires to provide no less than a unifying theory as a foundation for the analysis of industrial organization. I will perhaps be excused for feeling that this was an ambitious undertaking.

REFERENCES

Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.

Baumol, William J., Bailey, Elizabeth E., and Willig, Robert D., "Weak Invisible Hand Theorems on the Sustainability of Multiproduct Natural Monopoly," *American Economic Review*, June 1977, *67*, 350–65.

_____, **Panzar, John C., and Willig, Robert D.**, *Contestable Markets and the Theory of Industry Structure*, San Diego: Harcourt Brace Jovanovich, 1982.

Bertrand, Jules, Review of *Théorie Mathematique de la Richesse* and *Récherches sur les Principes Mathématiques de la théorie des Richesses*, *Journal des Savants*, 1883, 499–508.

Cournot, A. A., *Researches into the Mathematical Principles of the Theory of Wealth*, New York: A. M. Kelley, 1938; 1960.

Demsetz, Harold, "Why Regulate Utilities?," *Journal of Law and Economics*, April 1968, *11*, 55–65.

Fama, Eugene F. and Laffer, Arthur B., "The Number of Firms and Competition," *American Economic Review*, September 1972, *62*, 670–74.

Panzar, John C. and Willig, Robert D., "Free Entry and the Sustainability of Natural Monopoly," *Bell Journal of Economics*, Spring 1977, *8*, 1–22.

ten Raa, Thijs, "A Theory of Value and Industry Structure," unpublished doctoral dissertation, New York University, 1980.