~~1130A-80~~

Economies of Scale and Barriers to Entry

Richard Schmalensee

Sloan School of Management
Massachusetts Institute of Technology

1128-80

Dixit has recently presented a model in which established firms select capacity to discourage entry but cannot employ threats they would not rationally execute after entry. Entry deterrence in a slight modification of this model involves the classic limit-price output. Under linear or concave demand, however, the capital cost of a firm of minimum efficient scale is an upper bound on the present value of the monopoly profit stream that can be shielded from entry. It is argued that this suggests the general unimportance of entry barriers erected by scale economies.

Sloan WP # 1128-80

(Revised January 1981)

(*Journal of Political Economy*, forthcoming)

In his seminal work on conditions of entry, Bain (1956) argued that the necessity for a firm to be large relative to the market in order to attain productive efficiency created a barrier to entry. The notion that large-scale entry can create a discrete difference between pre-entry and post-entry price and profit levels is clear enough, but the details of Bain's argument are still somewhat controversial. In assessing the importance of the scale economies barrier, Bain introduced the limit-pricing model of entry deterrence, in which established firms act as a perfect cartel and potential entrants expect those firms to maintain their pre-entry levels of output even after entry. This model has been subjected to strong criticism, however, in large part because it may not be rational for the established firms to keep output constant after large-scale entry has occurred.[1] Moreover, Stigler (1968) and others have challenged the basic idea that scale economies can create a meaningful entry barrier. These critics have stressed the fact that once an entrant has invested in an efficient plant, there is no difference (under the usual assumptions) between its position and that of established firms. Without a post-entry difference, they have argued, there can be no real barrier to entry.

In part this last argument concerns the most useful definition of "barrier to entry". Spence (1977), building on insights of Schelling (1960), has recently presented an interesting analysis of the economic issues involved in that argument.[2] The basic point that emerges is that established firms, assuming they can coordinate their actions, have the advantage of being able to make some irreversible decisions before new entry appears. In particular, they can select the level of capacity facing any new entrant. Even if entry occurs and the established firms then wish to have less capacity, their pre-entry committments may make a rapid reduction in capacity impossible. Recognizing this, entrants may be deterred. In his formal analysis, Spence assumes that the established firms build enough capacity to produce nearly the competitive output. Before entry, they produce the monopoly output, but they threaten to use all their capacity if entry occurs. If this threat is believed, as Spence assumes it is, entry can clearly be deterred. But a threat to increase output after entry is surely no more credible than the threat to maintain output that is the core of the limit-price model.

In an important recent paper, Dixit (1980) explores the implications of restricting the established firms to threats that they would find it in their post-entry interests to execute. In particular, he assumes that potential entrants expect the post-entry market equilibrium to be that of Cournot duopoly. That is, they expect the established firms to behave rationally but noncooperatively in response to their entry. Though this market model is far from perfect, it at least implies plausible profit-seeking behavior by both established sellers and new entrants, and it avoids building ad hoc behavioral asymmetries into the post-entry situation.[3] Dixit finds that the established firms may still be able profitably to deter entry if they can commit to a level of capacity before potential entrants appear.

In Dixit's simplest model, scale economies arise because of fixed costs. In order to relate his analysis to the earlier limit-pricing literature and to empirical work on scale economies, it is assumed here instead that the long-run average cost curve is L-shaped, so that the minimum efficient scale of operations is well-defined.[4] The resultant model is analyzed in the next section. It is shown there that entry deterrence involves charging exactly the classic limit price before entry and holding output constant if entry occurs. A basic difference between this analysis and that of Bain-Sylos, however, is that the threat to hold output constant after entry is not always credible. It is shown that if the threat is credible, and if the market inverse demand function is linear or concave, the pre-entry present value of excess profits that can be shielded from entry cannot exceed the capital (startup) cost of a firm of minimum efficient scale. In Section II, it is argued that this upper bound and widely accepted estimates of the importance of scale economies imply that the entry barriers considered here are generally of little quantitative importance. The implications of localized competition and long-lived advertising in this context are discussed briefly.

I. Credible Entry Deterrence

Following most of the relevant literature, the established firms are assumed to be either a single enterprise or a perfect cartel and are referred to as "the monopoly" in what follows. Only one potential entrant, called "the entrant", is considered. The monopoly is assumed to select and construct its capacity, K^m, prior to the entrant's appearance. Both the monopoly and the entrant operate under the cost conditions shown in Figure 1. Economies of scale are such that firms with capacity less than K_0 have prohibitively high costs, so that K_0 is the minimum efficient scale in this industry. For levels of output, q, between K_0 and capacity, K, variable cost is v per unit. Capacity costs c per unit to install. It is assumed to last forever, so that the flow cost of capacity per period is (rc) per unit, where r is the relevant discount rate. With capacity K in place, the heavy kinked line in Figure 1 is the effective marginal cost schedule. By selecting K^m, the monopoly determines where the kink in its marginal cost function occurs. The entrant, with no sunk investment in capacity, faces a constant marginal cost of $t = rc + v$ for all output levels above K_0. All this follows the basic setup of Dixit's (1980) simplest model, except that he has $K_0 = 0$, and he allows non-zero fixed setup costs.

The first step in the analysis is the description of the monopoly's post-entry Cournot reaction function. For illustrative purposes, suppose that the industry inverse demand function is given by

$$P = P(Q) = a - bQ = a - b(q^e + q^m), \qquad (1)$$

where a and b are positive constants, P is price per unit, Q is total industry output, q^m is monopoly output, and q^e is the entrant's output. If marginal cost were some constant, X, the monopoly's reaction function would be simply

$$q^m = (a - bq^e - X)/2b. \qquad (2)$$

Things are more complex under the cost structure assumed here. The heavy kinked line in Figure 2 shows the monopoly's post-entry reaction function given installed capacity K^m and market demand given by (1).

In Figure 2, the line MM' is the reaction function of a firm chosing q^m, taking q^e as fixed, with constant marginal cost equal to t, while NN' is drawn for marginal cost v. If q^e exceeds U_v, the addition of K_0 to industry output would drive price below average variable cost, v, so that the monopoly's optimal output is zero even with capacity in place. For q^e between L_v and U_v, the monopoly optimally sets $q^m = K_0$. If there were no lower bound constraint on output, a smaller value of q^m would be chosen along NN'. As q^e is reduced below L_v, the monopoly expands output along NN' until the capacity constraint is encountered. The horizontal section of the reaction function in Figure 2 corresponds to the vertical portion of the marginal cost schedule in Figure 1. If q^e is sufficiently low, it may (depending on the initial choice of K^m) be optimal for the monopoly to add capacity and expand output, at a marginal cost of t per unit. As q^e goes to zero, the monopoly's optimal output goes to M, the unconstrained monopoly level, as Figure 2 is drawn. If K^m were chosen above M but below N, q^m would be set equal to K^m for $q^e = 0$. Monopoly output never exceeds N.

If the monopoly had not acquired any capacity, it would not invest if q^e were above U_t, since the addition of K_0 to industry output would drive the price below total unit cost, t. For q^e between L_t and U_t, it would be optimal to set $q^m = K_0$, while for smaller values of q^e it would be optimal to acquire more than the minimum possible capacity. Formally, the key quantity levels discussed above are given in general and in the linear case of equation (1) by

$$P(L_t+K_0) + K_0P_Q(L_t+K_0) = t; \quad L_t = (a-2bK_0-t)/b, \tag{3a}$$

$$P(U_t+K_0) = t; \quad U_t = (a-bK_0-t)/b, \tag{3b}$$

$$P(L_v+K_0) + K_0P_Q(L_v+K_0) = v; \quad L_v = (a-2bK_0-v)/b, \tag{3c}$$

$$P(U_v+K_0) = v; \quad U_v = (a-bK_0-v)/b. \tag{3d}$$

The reaction function relevant to the entrant's decisions is given by the mirror image of the schedule MAB in Figure 2. With no committment to capacity, the entrant has a constant (long-run) marginal cost of t for output levels above K_0. Consideration of Figure 2 then makes it clear that the entrant will be detered if it expects the monopoly's post-entry output to exceed U_t. From (3b), U_t is exactly the pre-entry output predicted by the classic Bain-Sylos limit-price theory, with $P(U_t)$ the corresponding pre-entry price.

Figure 3 depicts a situation in which entry is deterred by monopoly choice of a capacity level, K^m, above the limit quantity, U_t. (Note the reversal of axes between Figure 2 and Figure 3.) Given its precommittment to that capacity, the monopoly's reaction function is exactly as before; as long as q^e does not greatly exceed K_0, it is clear that the monopoly's optimal response is to utilize capacity fully. The entrant's reaction function, $\mu\alpha\beta$, is the mirror image of the ex ante monopoly schedule MAB. Note that with the monopoly's capacity in place and the corresponding costs sunk, the conventional Cournot equilibrium point, e, has no particular significance. The only equilibrium is $q^m = K^m$, $q^e = 0$; entry is deterred, and the monopoly always uses its installed capacity fully.

A bit more analysis of the situation depicted in Figure 3 makes it clear that in order for the monopoly to be able to deter entry and still enjoy positive economic profit, two conditions are necessary. First, c must be positive; it must be possible for the monopoly to commit itself in advance to capacity, to purchase a lower marginal cost over some output range. If c = 0, the MC = v reaction schedule N"N (which corresponds to the left-hand portion of NN' in Figure 2) collapses to M'M, and the monopoly cannot avoid the Cournot point, e, which involves entry. Second, K_0 must be positive; there must be scale economies as well as durable capital. If $K_0 = 0$, the entrant's reaction function extends to the point μ', and K^m must exceed μ' in order to deter entry. But an examination of the entrant's reaction function makes it clear that $P(\mu') = t$, so that the monopoly can deter entry only by producing at least the competitive output in the absence of scale economies. (When $q^e = 0$, the entrant's marginal revenue is exactly the market price. Reference to (3b) makes it clear that $U_t < \mu'$, as Figure 3 shows.)

A somewhat stronger necessary condition for entry deterrence with positive profits is that the point β in Figure 3 must lie inside the $N''N$ locus. That is, if marginal cost is v, it must be optimal for the monopoly to produce at least U_t in response to $q^e = K_0$. If not, any entrant can plan on producing K_0 and earning positive profit, regardless of the monopoly's choice of capacity or pre-entry output. If the market marginal revenue curve slopes downward, this condition means that the monopoly's marginal revenue at the point β must be at least v:

$$P(U_t+K_0) + U_tP_Q(U_t+K_0) \geq v. \tag{4}$$

Rearranging, noting that $P(U_t+K_0) = t$ from (3b), and multiplying by K_0, one obtains

$$[-P_Q(U_t+K_0)K_0]U_t \leq rcK_0. \tag{5}$$

If the inverse demand function is either linear or concave, weak concavity implies

$$P(U_t) \leq P(U_t+K_0) + P_Q(U_t+K_0)[U_t - (U_t+K_0)]. \tag{6}$$

Substitution from (6) for the bracketed term in (5) yields

$$[P(U_t) - t]U_t \leq rcK_0. \qquad (7)$$

Since U_t is the monopolist's minimum (pre-entry and post-entry) deterrence output, condition (7) says that the flow of excess profit enjoyed by a monopoly that has credibly deterred entry cannot exceed the per-period cost of capital employed in an enterprise of minimum efficient scale.[5] This inequality clearly reflects the fact that both c and K_0 must be positive for this sort of limit-pricing to be credible and profitable.

It should be noted that it may not be optimal for the monopoly to install sufficient capacity to deter entry even if it is possible to do so. Depending on the details of the demand structure, the monopoly might prefer to allow entry and have the equilibrium occur at some point along $e\alpha\beta$. It can, of course, guarantee itself any such point within N'N by its choice of K^m. If the monopoly does decide to admit the entrant and the latter invests in capacity, inequality (7) serves to bound the total excess profit that the resultant duopoly can protect from entry by coordinated action.[6] In general, if entry is unattractive to outsiders, the flow of monopoly profits received by cooperating insiders cannot exceed the flow cost of capital assets embodied in a firm of minimum efficient scale, as long as industry demand is not strictly convex.

II. Extensions and Implications

The derivation of (7) rests on a number of strong assumptions that serve to overstate the ease of deterrence. First, it is at least arguable that the Cournot model overstates the intensity of rivalry to be expected in the post-entry duopoly.[7] If the entrant is more optimistic, entry will be harder to deter, and the flow of excess profits that can be protected will be smaller than (7) allows. Second, it was assumed that the monopoly was able to invest in infinitely durable assets. The recent work of Eaton and Lipsey (1980) supports the intuitive argument that the strength of the monopoly's commitment is reduced as the assets it purchases become less durable, since the monopoly cannot generally commit itself in advance to invest after entry in order to maintain capacity. Third, the assumption that average cost is effectively infinite for outputs below K_0 is very unrealistic. As the presentations of Modigliani (1958) and Scherer (1980, ch. 8) show, entry deterrence in the Bain-Sylos limit pricing model is more difficult the more slowly average cost rises when output is reduced below K_0. The need to deter entry at small scale lowers the profitability of the limit-price point, which is exactly the point considered in deriving inequality (7). Finally, the assumption that the established firms manage to coordinate their investment policies perfectly in the interest of entry deterrence is very strong indeed. Less than perfect coordination in deterrence would be expected to make entry more likely.

Given the strength of these four assumptions, it would seem sensible to use (7) as an upper bound on monopoly profit associated with scale economy entry barriers even in situations where industry inverse demand may be convex. To explore the implications of this conclusion, let us consider an industry with (price level adjusted, net) assets A, earning an accounting rate of return r^* on those assets. Assuming away measurement problems and windfalls, accounting profits equal normal returns to capital plus monopoly profits. Let the competitive rate of return, the ratio of normal profits to A, be r. It is then sensible to use r to go from stocks to flows in this industry as is done in (7). That condition can then be re-written as

$$(r^* - r)A \leq rcK_0, \qquad (7')$$

which immediately becomes

$$(r^* - r)/r \leq cK_0/A. \qquad (8)$$

Assuming that economies of scale are essentially exhausted at capacity K_0, the ratio cK_0/A should be a good approximation to the ratio K_0/Q, the ratio of minimum efficient scale (in output terms) to market demand. For most industries, most scholars estimate this ratio to be less than 0.10. (Scherer (1980, ch. 4) presents a large number of such estimates.) Condition (8) then implies that if an industry has a return on assets of, say, 10%, entry is not taking place, and only scale economies can serve to deter entry, the normal rate of return must be at least 9.1% unless scale economies are exceptionally important. Scale economy entry barriers thus cannot account for a rate of monopoly profit of even 1%, surely not much above than the standard deviation of the corresponding measurement error in most situations.

An analysis of deadweight loss similarly suggests that entry barriers derived from scale economies are not generally important. If industry demand is given by (1), it is easy to show that the deadweight loss from limit pricing is $b(K_0)^2/2$. Letting demand elasticity, P/bQ, equal E, the ratio of deadweight loss to sales revenue becomes $[(K_0/Q)^2/2E]$.

This exceeds 0.005 only if economies of scale are of exceptional importance or if E is less than one.

The analysis so far, based on conventional wisdom about the importance of scale economies, supports Bain's (1956, p. 212) conclusion that "economies of scale in production and distribution do not loom large as the basis of barriers to entry." Two factors that might weaken this conclusion deserve brief mention, however.

First, most U.S. estimates of the importance of scale economies assume national markets and implicitly treat products as homogeneous. If transport costs are important and markets are regional, scale economies can become very important indeed. Scherer (1980, p. 98) estimates that a cement plant of minimum efficient scale would account for about 40% of demand in a typical regional market, for instance. If products are differentiated, competition may be similarly localized in product space. As is discussed at length in Schmalensee (1978), this sort of demand structure also serves to magnify the importance of scale economies in entry deterrence.

Second, spending on advertising generally has an investment component, though there is considerable uncertainty about the typical durability of the corresponding asset.[8] The analysis of Section I above implies that by itself, the longevity of advertising's effects on demand does not make it possible to use advertising to protect excess profits.[9] A number of authors have suggested that the use of advertising involves significant economies of scale, however. There is also considerable uncertainty about the importance of these scale effects.[10] If it turns out in general or in some industry that scale economies in advertising are important <u>and</u> that advertising has relatively long-lasting effects on demand, the analysis above would lead one to suspect, by analogy, that advertising could be used like investment in plant and equipment to protect excess profits from entry. In order to verify this suspicion, however, one would at least have to verify that at any instant past investment in advertising has the same sort of impact on <u>marginal</u> returns to changes in output that investment in capacity does. The first of the models discussed in Schmalensee (1974) illustrates that advertising does not have such effects in at least some plausible dynamic models.[11] Rather than attempting to reason by analogy from the treatment of investment in capacity here, it would seem preferable to construct explicit models involving economies of scale in advertising and durability of advertising's effects on demand and to see if those models allow established monopolies to deter entry by means of investment in advertising. Such models may even imply simple analogs of (7), but that remains to be seen.

References

Bain, J.S. Barriers to New Competition. Cambridge, Mass.: Harvard University Press, 1956.

Comanor, W.S., and Wilson, T. "The Effect of Advertising on Competition: A Survey," Journal of Economic Literature 17 (June 1979): 453-76.

Dixit, A. "A Model of Duopoly Suggesting a Theory of Entry Barriers," Bell Journal of Economics 9 (Spring 1979): 20-32.

_______. "The Role of Investment in Entry-Deterrence," Economic Journal 90 (March 1980): 95-106.

Eaton, B.C., and Lipsey, R.G. "Exit Barriers are Entry Barriers: The Durability of Capital as a Barrier to Entry," Bell Journal of Economics, 11 (Autumn 1980): 721-29.

Little, J.D.C. "Aggregate Advertising Models: The State of the Art," Operations Research, 27 (July-August 1979): 629-67.

Modigliani, F. "New Developments on the Oligopoly Front," Journal of Political Economy 66 (June 1958): 215-32.

Orr, D., and MacAvoy, P.W. "Price Strategies to Promote Cartel Stability," Economica 32 (May 1965): 186-97.

Salop, S.C. "Strategic Entry Deterrence," American Economic Review 69 (May 1979): 335-8.

Schelling, T.C. The Strategy of Conflict. Cambridge, Mass.: Harvard University Press, 1960.

Scherer, F.M. Industrial Market Structure and Economic Performance, 2nd Ed. Chicago: Rand McNally, 1980.

Schmalensee, R. The Economics of Advertising. Amsterdam: North-Holland, 1972.

Schmalensee, R. "Brand Loyalty and Barriers to Entry," *Southern Economic Journal* 40 (April 1974): 579-88.

__________. "Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry," *Bell Journal of Economics* 9 (Autumn 1978): 305-27.

Spence, M. "Entry, Capacity, Investment and Oligopolistic Pricing," *Bell Journal of Economics* 8 (Autumn 1977): 534-44.

_________. "Investment Strategy and Growth in a New Market," *Bell Journal of Economics* 9 (Spring 1979): 1-19.

_________. "Notes on Advertising, Economies of Scale, and Entry Barriers," *Quarterly Journal of Economics*, 95 (November 1980): 493-508.

Stigler, G.J. "Barriers to Entry, Economies of Scale, and Firm Size," In G.J. Stigler, *Industrial Organization.* Homewood, Ill.: R.D. Irwin, 1968.

Footnotes

I am indebted to Steve Salop, Roger Bohn, and Nalin Kulatilaka for helpful comments, though I cannot share blame for remaining defects with them.

1. Scherer (1980, ch. 8) provides a useful discussion of the limit-pricing model and its critics.
2. See also Dixit (1979), Spence (1979), and, for a good discussion of the basic concepts involved, Salop (1979).
3. In addition to the discussion of the Cournot assumption by Dixit (1980), see Orr and MacAvoy (1965), where it is argued that Cournot behavior is a plausible threatened reaction to cheating on a cartel agreement.
4. See, for instance, Modigliani (1958) and Scherer (1980, chs. 4 and 8).
5. In this model, there would seem to be no general way to obtain a bound on profit like condition (7) without using concavity. (If demand elasticity is a constant, $-E < -1$, (4) is satisfied for all positive K_0 if and only if $v \leq rc(E-1)$.) Similar bounds hold in some related models of entry deterrence. In the "simple illustrative formal model" of Schmalensee (1978, pp. 312-3), for instance, if new brands don't expand total demand, per-brand excess profit in deterrence equilibrium equals brand-specific fixed cost. With output expansion, fixed cost becomes an upper bound. See also the equilibria in the references cited in Schmalensee (1978, p. 313), footnote 15.

6. If a given level of capacity is divided among several firms, each will see a larger Cournot marginal revenue at capacity output than would a single seller. For any given entrant output, capacity is thus more likely to be fully utilized with several established firms than with a single monopolist. This means that a higher level of profits can in

principle be protected against entry by Cournot behavior than the monopoly analysis in the text implies. Given the difficulty of coordinating investment policies among several firms, however, there seems little to be gained by a detailed analysis. It is surely reasonable, though not necessarily correct, to assume that monopoly entry deterrence involves higher pre-entry profits than feasible oligopoly deterrence strategies.

7. See the discussion of that model in Scherer (1980, ch. 5).

8. Comanor and Wilson (1979) survey the evidence on this point.

9. This same argument is made in Schmalensee (1974), though in a rather different framework. See also Spence (1977).

10. See Schmalensee (1972, ch. 7) and Comanor and Wilson (1979). Spence (1980) provides a useful discussion of the definition of economies of scale in this context, though he does not treat advertising as an investment.

11. But see also the "alternative model" discussed briefly in Schmalensee (1974, pp. 586-7) and the treatment of advertising in Spence (1977). To the extent that a certain minimum amount of introductory advertising is necessary to make any sales at all, such introductory outlays would seem to have the required impact on marginal returns. In Schmalensee (1978), such introductory advertising is treated as a lumpy investment central to entry deterrence. Little (1979) provides an illuminating survey of dynamic advertising/sales models.

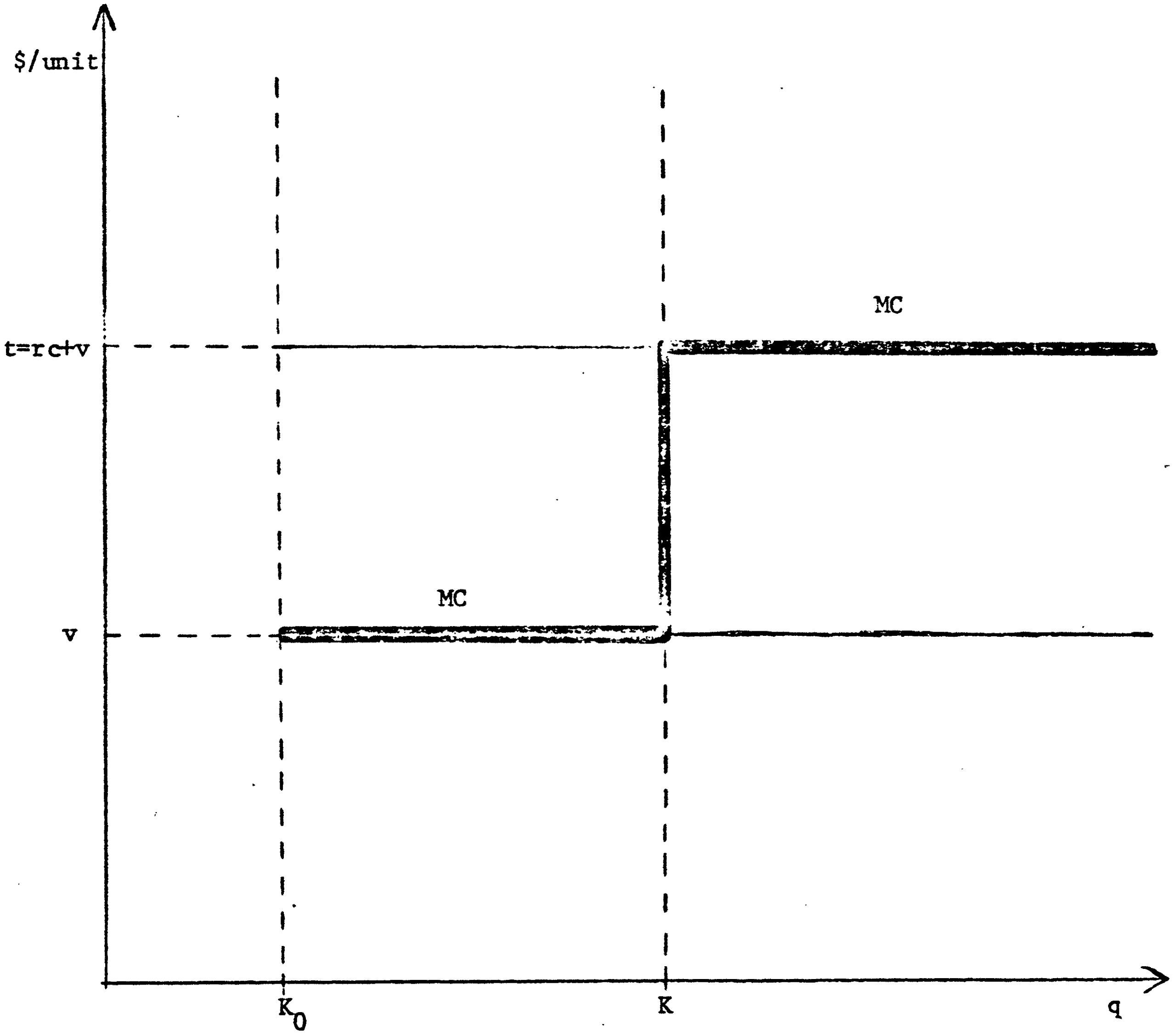

Fig. 1 Marginal Cost After Acquisition of Capacity K

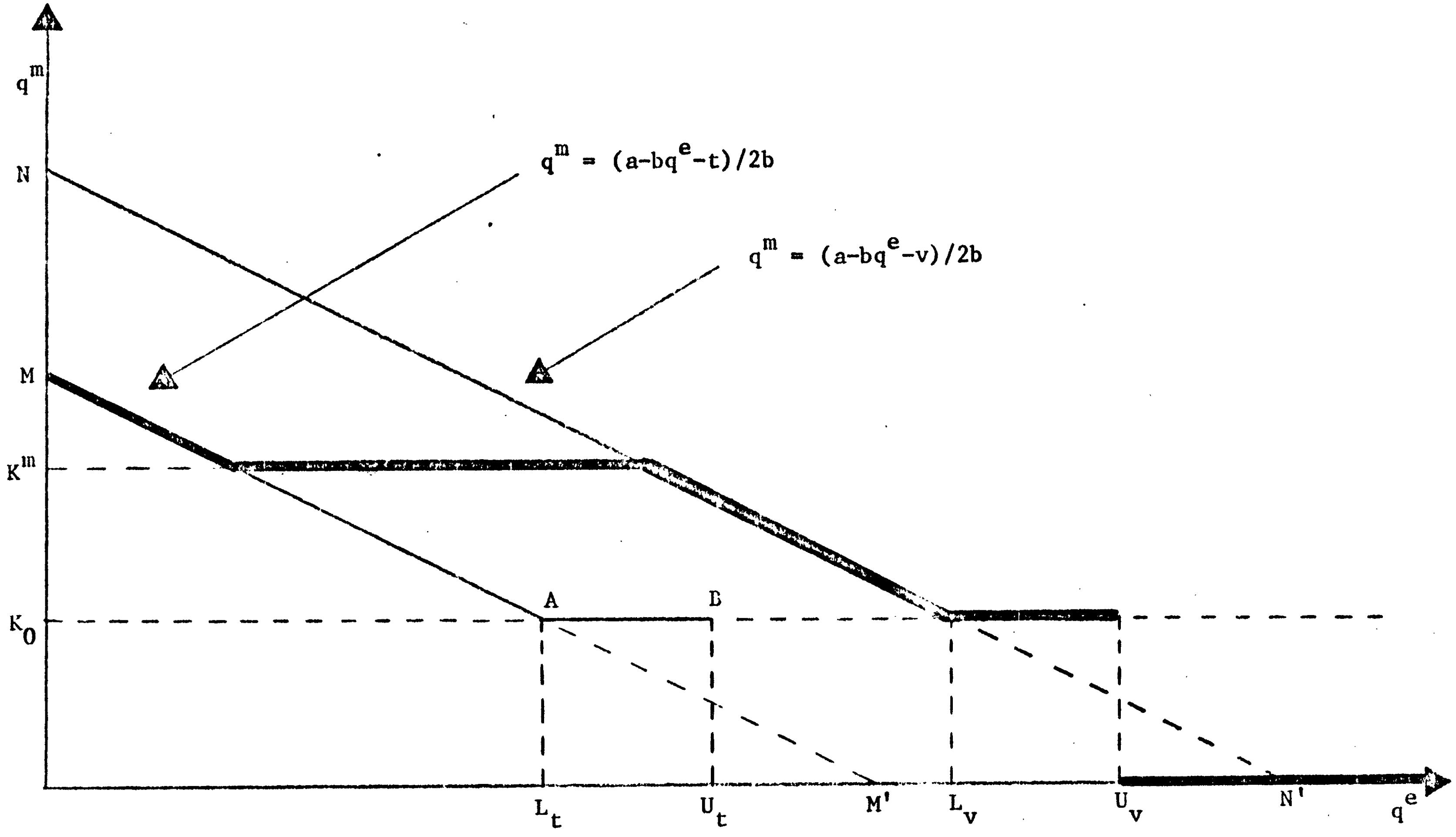

Fig. 2 The Monopoly's Post-Entry Reaction Function

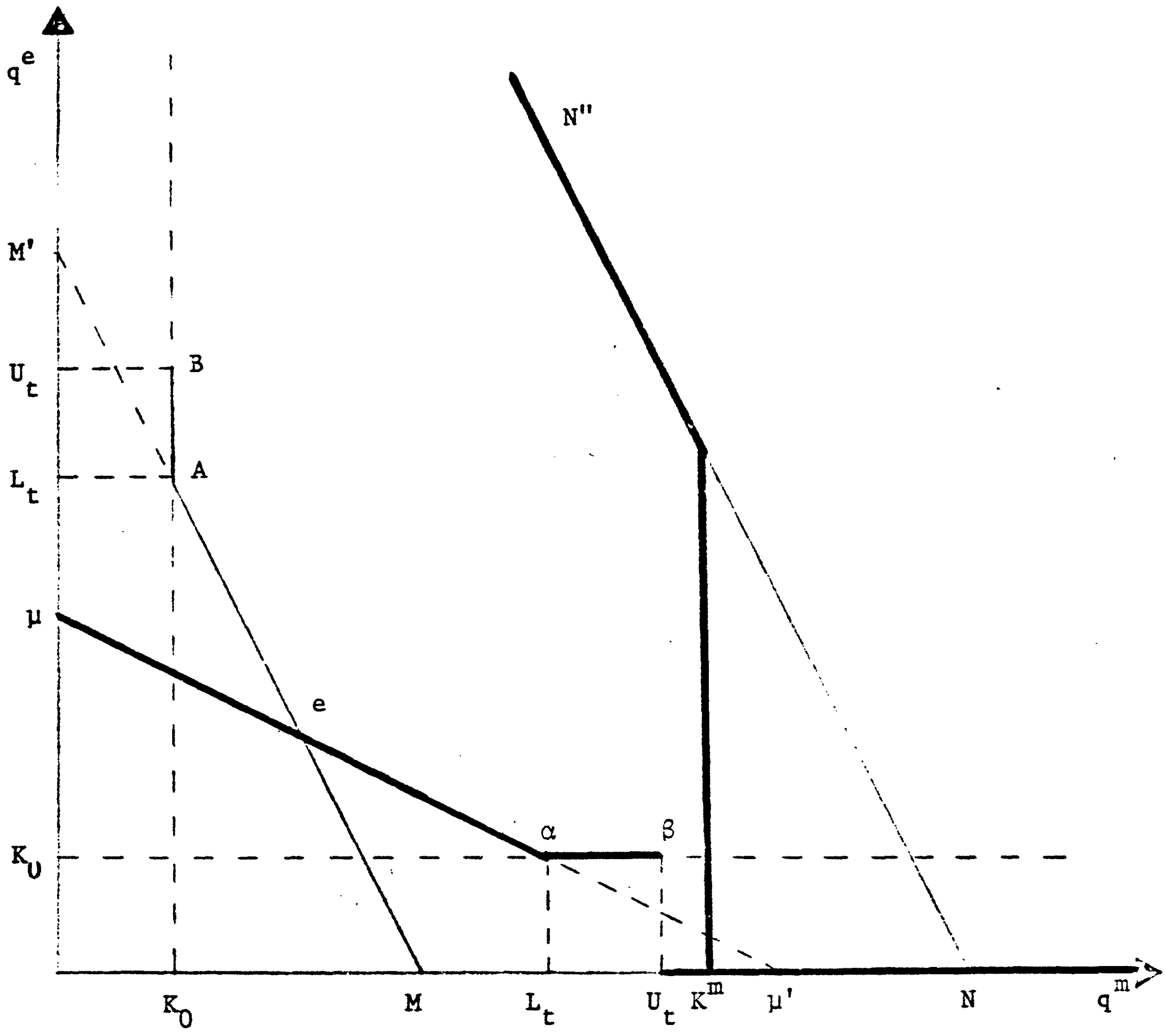

Fig. 3 Credible Entry Deterrence