Entry, capacity, investment and oligopolistic pricing

A. Michael Spence
Professor of Economics
Harvard University

The paper argues that entry is deterred in an industry when existing firms have enough capacity to make a new entrant unprofitable. This capacity need not be fully utilized in the absence of entry. This can result in larger costs than are necessary, given output levels. It also results in higher prices and lower levels of output than those implied by various forms of the limit price model. Capacity and other forms of investment are effective entry deterring variables, partly because they are irreversible and represent preemptive commitments to the industry.

1. Introduction

The purposes of this paper are two. One is to argue that entry in an industry selling a relatively homogeneous product is likely to be affected by the relation between demand and industry capacity and to trace the implications for market performance of this approach to entry. The second is to suggest that, in general, entry can be deterred by investment decisions (one of which is capacity) and that the investment approach to entry allows for a unified treatment of entry in the homogeneous and differentiated product cases.

The basic idea developed below is that an industry will carry excess capacity to deter entry. As a result, price will exceed the limit price, and production will be inefficient. The threat of entry effectively places a lower bound on capacity or capital. In this respect, the threat of entry is similar in its effects to a constraint on the industry’s rate of return. The height of the structural entry barriers determines the stringency of the constraint. The paper concludes with a simple probabilistic model that is applicable to differentiated product industries, where capacity may be a less important entry-deterring investment than advertising or a system of exclusive dealerships.

2. Model 1: capacity as an upper bound on output

The principle of this model is quite simple. It is that existing firms choose capacity in a strategic way designed to discourage entry. This strategic purpose is realized by holding “excess” capacity in the preentry period. This excess capacity permits existing firms to expand output and reduce price when entry is threatened, thereby reducing the prospective profits of the new entrant who operates on the re-

This work was supported by National Science Foundation Grant GS-39004. I am indebted to Kenneth Basemann, Ronald Braeutigam, Richard Caves, Michael Hunt, Bruce Owen, David Starrett, and Robert Wilson for their comments.
sidual demand curve to zero. Given that capacity is selected in this entry forestalling manner, existing firms choose preentry price and quantity so as to maximize profits.

Notation. Capacity, measured in output units is $k$. The annual cost of capacity (interest on debt, or opportunity cost if funds are raised internally) is $r$. Variable costs are $c(x,k)$. In this first model it is assumed that $c_{xk} = 0$: capacity has no effect on marginal costs. The inverse industry demand function is $P(x)$, and quantity is $x$. Revenue is $R(x) = xP(x)$. Profits, denoted by $\pi(x,k)$, are

$$\pi(x,k) = R(x) - c(x) - rk.$$ 

Average total costs are denoted by $a(x,k) = (c(x)/x) + (rk/x)$. When $k = x$, the firm or industry is efficient in that costs are minimized, given the level of output $x$. In this case, $a(x,x) = (c(x)/x) + r$.

The industry with no entry threat. With no threat of entry, the industry maximizes $\pi(x,k)$ with respect to $x$ and $k$, subject to $x \leq k$. It is clear that $x = k$ at an optimum (because $k$ does not affect marginal costs by assumption). Therefore, at a maximum,

$$R'(x) = c'(x) + r,$$

and the capacity constraint is binding.

Entry threats. Suppose the existing industry sets capacity at $k$. Then when entry begins to occur, it can expand output to $k$ and lower price to $P(k)$ within the time horizon required for entry. With respect to the time required for the installation of capacity, the industry is on an equal footing with the potential entrant. It is assumed that the existing industry can hold the output $x = k$ in the face of entry. The demand left over to the entrant (often called the residual demand) is defined as follows. If the entrant supplies $y$, total industry output is $k + y$. Therefore, price is $P(k + y)$, and this, regarded as a function of $y$, is the inverse demand function facing the entrant. Entry is deterred if for all $y$, profits for the entrant are nonpositive: that is to say, for all $y$

$$P(k + y) \leq a(y,y) = \frac{c(y)}{y} + r.$$ 

Equation (1) simply says price falls short of average cost for all quantities. It is clear that as $k$ increases, $P(k+y)$ falls for each $y$. Indeed, for $k$ sufficiently large, residual demand is zero. Therefore, there is a minimum level of $k$, denoted $\tilde{k}$, for which (1) holds. If the existing industry maintains capacity $\tilde{k}$, entry is deterred.

The existing industry maximizes profits with respect to $x$ and $k$, subject to two constraints:

$$x \leq k,$$

which indicates that quantity does not exceed capacity, and

$$k \geq \tilde{k},$$

Section 3 relaxes this assumption.

This is the assumption of the limit price theory. It can be relaxed probabilistically: see Kamien and Schwartz (1971) and subsequent sections of this paper.
which ensures that entry is deterred. The Kuhn-Tucker conditions for
this problem are (with multipliers of \( \lambda \) and \( \mu \))

\[
\begin{align*}
R'(x) - c'(x) &= \lambda \\
\mu &= \lambda + \mu \\
\lambda(k - x) &= 0, \\
\mu(k - \bar{k}) &= 0 \\
\lambda, \mu &\geq 0.
\end{align*}
\]

Case 1: \( \mu = 0, \lambda = r \). If \( \mu = 0 \), then \( \lambda = r \) and \( R' = c' + r \). This
occurs when the unconstrained profit maximizing decisions establish
capacity at a level at which entry is automatically deterred. This will
occur when demand is highly inelastic in the range of prices near
marginal cost.

Case 2: \( \lambda = 0, \mu = r \). If \( \lambda = 0 \) and \( \mu = r \), the industry sets \( k = \bar{k} \) and
then maximizes profits by setting \( x < \bar{k} \) and price above \( P(\bar{k}) \). The
constraint \( x \leq k \) is not binding. More importantly, costs are not
minimized, given the level of output actually supplied. For the pur-
pose of deterring entry, capacity is maintained above the efficient
level, given output.

Case 3: \( \lambda \neq 0, \mu \neq 0 \). If both multipliers are nonzero, then both
constraints are binding: \( k = \bar{k} \) to deter entry, and given that capacity
is at the level \( \bar{k} \), \( x = \bar{k} \) to maximize profits. Unlike Case 2, if the threat
of entry were removed here, capacity and output would come down
together.

Figure 1 illustrates Cases 2 and 3, these being of greatest interest.
It is to be noted that with the threat of entry present, the overall
industry cost curve \( C(x) \), is as follows:

\[
C(x) = c(x) + rk, \quad x \leq \bar{k} \\
= r, \quad x > \bar{k}.
\]

In each case, the heavy line is the industry cost function in the
presence of a threat of entry. The quantity \( x^* \) is the industry’s optimal
output without an entry threat. In Case 2, the entry-constrained
optimum is at \( \bar{x} < \bar{k} \), while in Case 3, the constrained optimum is at
\( \bar{x} = \bar{k} \). Essentially, once capacity \( \bar{k} \) is installed to deter entry, it may
or may not be profitable to expand output to \( \bar{k} \). It depends on the
structure of demand.

\( \square \) Comparison with the limit price theory. Since the limit price theory
is thoroughly treated in the literature, I shall outline it only briefly to
facilitate comparison. It maintains that low price and high quantity
are the proximate deterrents of entry. If \( \bar{x} \) is industry output, then
residual demand is \( P(\bar{x} + y) \). Entry is deterred if \( \bar{x} \) is sufficiently large
to leave no profits in residual demand. Hence the industry will either
set \( x \) greater than the minimum \( \bar{x} \), required to deter entry (the
analogue of Case 1), or set \( x = \bar{x} = \bar{k} \) and \( \bar{p} = P(\bar{x}) \). This is Case 3.
Note that production is always efficient. There is never excess capac-
ity, given the level of output. What is ruled out by the limit price
approach is Case 2, where \( x < \bar{k} \).

The theories differ in that the excess capacity hypothesis implies
that price may exceed the limit price, and quantity be lower than the
limit quantity. When this occurs, the industry carries excess capacity,
given output, and production is inefficient. To the extent that entry
affects price, it is through the impact of capacity on the structure of
the cost function. In some cases, the profit maximizing price when the
industry has sufficient capacity to deter entry may be the limit price,
but it may not.

The limit price theory suggests that the short-run oligopoly pricing
game cannot be conducted independent of the threat of entry. If limit
pricing is accepted, analyses of the pricing game that ignore entry
make little sense. The capacity hypothesis suggests somewhat the
reverse; that the short-run pricing game is to some extent, strategi-
cally independent of entry. It is not totally independent because of the
aforementioned effect of capacity on costs.

**Capacity, price and market share.** Consider the industry without
the threat of entry. It will maximize profits by setting $R' = c' + r$ and
$x = k$. Usually, when price exceeds marginal cost, there is an incentive
for firms in the industry to shade or cut price and expand output.
However, with capacity at $k = x$, there is no incentive to do so. In
fact, tacitly agreeing on capacity may be a powerful technique for
enforcing tacit collusion on price. In this respect, capacity has the
distinct advantage of being difficult to expand without being noticed.
Now the effect of entry is to remove the capacity level as a means of
enforcing a supercompetitive price level in the industry. Therefore,
under some conditions, the threat of entry may operate on price by
making it more difficult to maintain price at a high level. In Case 2, at
least one firm in the industry is operating with excess capacity, and
that firm has an incentive to cut price and expand output. The threat
of entry forces the industry to find other means of agreeing on market
share than by tacitly agreeing to hold capacity down.  

---

3 When capacity affects marginal costs as in the next section, holding capacity down
to remove price cutting incentives does not give rise to a situation in which industry
profits are maximized. But when capacity is a pure upper bound on output, the industry
profit maximizing position is achievable.
3. Model 2: capacity affects marginal costs

If capacity is interpreted as irreversible capital investment in plant and equipment, then it may serve to set an upper bound on output, but it is also likely to affect the marginal costs in the firm and the industry. The choice of capacity in the industry is really the choice of the cost function with which it will operate in the short run and in response to entry. Formally, variable costs \( c(x,k) \) depend on \( k \). It is assumed that \( c_{xk} < 0 \); increases in capacity reduce the marginal variable cost. Total costs are \( c(x,k) + rk \). We define the minimized costs for each level of output as follows:

\[
S(x) = \min_k [c(x,k) + rk].
\]

Costs are minimized, given \( x \), when \( c_k + r = 0 \).

The purpose of this section is to extend the analysis of entry to the case of an industry with costs of this form. It is clear that the previous case is a limiting version of the present one in which marginal variable costs are unaffected for \( x \leq k \), and at \( k \) they become infinite.

Suppose an entrant selects capacity \( k \) and enters. Pricing discipline in the industry breaks down (as it did in the previous model), and price equals marginal cost. The equilibrium price \( p \), and quantities, \( \tilde{x} \) for the existing industry and \( \tilde{y} \) for the entrant, are therefore given by

\[
\tilde{p} = p(\tilde{x} + \tilde{y}) = c_x(\tilde{x},k) = c_x(\tilde{y},k). \tag{2}
\]

At \( (\tilde{y},\tilde{k}) \) average costs for the entrant are \( (c(\tilde{y},\tilde{k})/\tilde{y}) + r\tilde{k} \). If these exceed \( \tilde{p} \), then his profits will be negative. Notice that \( \tilde{p}, \tilde{y}, \) and \( \tilde{x} \) depend on \( \tilde{k} \) and \( k \) from (2). Entry is deterred by the existing industry's capacity \( k \), if for all \( \tilde{k} \),

\[
\tilde{p}(\tilde{k},\tilde{k}) \leq c(\tilde{y}(\tilde{k},\tilde{k}),\tilde{k}) + r\tilde{k}.
\]

Let \( \tilde{k} \) be the smallest level of \( k \) that ensures this outcome. Then \( \tilde{k} \) is the level of capacity in the existing industry required to deter entry. It is assumed that the industry sets capacity at or above this level.

The preceding analysis can be illustrated diagrammatically in terms of residual demand. Once \( k \) is selected, the existing industry has a well-defined marginal cost curve, \( c_x(x,k) \). It is depicted in Figure 2.

When entry occurs, the existing industry will supply along the curve \( c_x(x,k) \). This leaves as residual demand the horizontal distance between the marginal cost curve and the demand curve, \( ABC \). For example, at price \( \tilde{p} \), the existing industry supplies \( x^* \), leaving \( \tilde{x} - x^* \) for the potential entrant. Entry is deterred if the potential entrant cannot make a profit in the residual demand. Notice that if \( k \) is increased, the marginal cost curve moves to the right (the dotted line in Figure 2) and residual demand is reduced.

The existing industry solves the following problem:

\[
\max_{\tilde{x}} P(\tilde{x}) - c(\tilde{x},k) - r\tilde{k}
\]

subject to \( k \geq \tilde{k} \).

---

4 At the Nash equilibrium in an oligopoly selling a homogeneous product, with both price and quantity as strategy variables for the firms, price equals marginal cost.
5 If the industry launches an all out attack, and is prepared to supply at zero profit, it will move along \( a(x,k) \), and residual demand is \( DBC \).
The Kuhn-Tucker conditions are

\[ R'(x) - c_x(x,k) = 0, \]

and

\[-c_k - r + \lambda = 0,\]

with

\[\lambda[k - \bar{k}] = 0\]
\[\lambda \geq 0.\]

If \(\lambda = 0\), then the unconstrained profit maximizing capacity is greater than \(\bar{k}\), and entry is deterred automatically. This is the analogue of Case 1 in the previous section. Otherwise \(\lambda > 0\), the constraint is binding, and

\[c_k + r = \lambda > 0,\]

implying that capacity is higher than optimal, given the level of output in the industry. That is to say, production is inefficient (i.e., costs are not minimized), given the level of output.

From the condition, \(R'(x) = c_x(x,k)\), we have upon differentiation\(^6\)

\[ \frac{dx}{dk} = \frac{c_{xk}}{R'' - c_{xx}} > 0.\]

Therefore, the profit maximizing quantity, given \(k\), increases with \(k\). Since \(k = \bar{k}\) is higher than the unconstrained profit maximizing level of \(k\), it follows that output will be higher and price lower than in the unconstrained case. Thus the threat of entry does affect price, via the effect of capacity on marginal costs.

\(\square\) **Comparison with the limit price.** In the limit price theory, entry is

\(^6\) The denominator is negative because of the second-order condition for a maximum in \(x\), \(R'' - c_{xx} < 0.\)
deterred by price. Consider Figure 3. If the industry selects \((\hat{p}, \hat{x})\) for price and quantity, residual demand is the triangle \(ABD\). Capacity is set to minimize costs. The average and marginal cost curves are shown. Now if \(ABD\) is sufficient to deter entry, then residual demand \(MBD\) under the capacity theory, at the limit price capacity, is more than sufficient to deter entry, for in the excess capacity theory, the potential entrant anticipates some quantity increases by the existing industry. Therefore, capacity will be larger under the limit price theory than the capacity theory. In other words, the industry operates with less capacity than the limit price theory suggests.\(^7\)

But there is a further effect. Even if capacities were the same under the two hypotheses, output would be less and price higher under the capacity hypothesis, because price, not being required to deter entry, can rise. Thus, price is higher and quantity lower under the capacity hypotheses for three reasons: (1) entry deterring capacity is less than limit price capacity, (2) profit maximizing quantity rises as capacity rises and falls as capacity falls, and (3) profit maximizing quantity at the limit price capacity is lower than the limit quantity.

Somewhat more simply, if we started an industry at the limit price, quantity, and capacity, it would lower output to maximize profits, then lower capacity and lower output more. In short, the excess capacity theory predicts price is higher and quantity and capacity lower than the limit price theory. Moreover, capacity is above the optimum given output, and the short-run oligopoly pricing game is relatively independent of entry threats. It is affected only through downward pressure exerted by entry deterring capacity on costs and hence, constrained profit maximizing price.

To summarize the static analysis thus far, if entry is deterred by excess capacity, because that partially determines the prospects for the entrant once prices and quantities shift, and the entrant knows this, the oligopolistic pricing will not be so low as the limit price.

\(^7\) If capacity is literally an upper bound on output, the capacities under the two theories would be the same.
theory suggests, and costs will not be minimized. It is the latter that provides the basis for testing the model against the leading competitor. That is to say, the presence of capacity above cost minimizing levels, given output, would be evidence for the capacity theory and against the limit price theory. In a dynamic context, capacity running well ahead of demand would be observed.

If we take the sum of consumer and producer surplus as a measure of market performance, the threat of entry enlarges the consumer surplus by forcing capacity, and hence quantity, up. It also reduces profits or producer surplus by moving the industry off the profit maximizing point. One can ask about the total or overall effect. If production were efficient, given output, the total surplus would rise, because the loss to the producers would be a pure transfer to consumers. This is what the limit price theory suggests. But when capacity is expanded to deter entry, total surplus can be lower with a threat of entry than without it. And it will certainly be lower than it is at the limit price with efficient production.

The threats of entry and rates of return. It is of some interest to note that the threat of entry implicitly constrains the rate of return. As we have seen, the industry maximizes

\[ \pi(x,k) \]

subject to \( k \leq \tilde{k} \).

Let

\[ M(k) = \max_x \pi(x,k). \]

The industry's return on capital is \( \frac{M(k)}{k} + r \). As shown in Figure 4,

![Figure 4: Capacity and ROR Constraints](image)

the rate of return rises and then falls. The effect of constraining \( k \) to be greater than \( \tilde{k} \) is to constrain the rate of return to be less than \( s = \frac{M(\tilde{k})}{\tilde{k}} + r \). That is, the threat of entry, in conjunction with the structural entry barriers, constrains the rate of return on capital.\(^8\)

\(^8\) In some respects, the effects of the threat of entry discussed earlier are similar to the effects of a rate of return constraint for a regulated monopoly. (See Averch and
The previous analysis of capacity and entry is formally applicable only to homogeneous product industries with some economies of scale. However, the spirit of the analysis carries over to more general cases.

The central feature of the entry deterring strategy is that the industry invests now in some form of capital, and that that investment reduces the prospects of a potential entrant and hence the probability of entry. The investment can be plant and equipment and the capital capacity. But the investment could also be in advertising and the capital customer “loyalty.” The investment could also be the installation of retail outlets and the capital the distribution system.

A simple model will serve to illustrate the points. Let \( q \) be price and \( z \) the level of some stock. It could be capacity, or a weighted sum of present and past advertising expenditures. Profits in the current period are \( \pi(q,z) \). The probability that no entry will occur is an increasing function of \( z \), denoted \( p(z) \). Let \( V^* \) be the expected present value of the profit stream if entry does occur, the present value being with respect to the next period. \(^9\) \( V(q,z) \) is the expected present value of industry profits. The discount rate is \( d \), as before.

The tree in Figure 5 illustrates the decision problem facing the firm or industry. The function \( V(q,z) \), the present value of profits, is defined by

\[
V(q,z) = \pi(q,z) + p(z) \frac{V(q,z)}{1 + d} + (1 - p(z)) \frac{V^*}{1 + d}.
\]

\(^9\) I shall take \( V^* \) to be a number that has already been estimated for the purposes of this model.
Profits are maximized at values of \( q \) and \( z \) which satisfy the first-order conditions

\[
\pi_q = 0, \quad (i)
\]

and

\[
\pi_z + \frac{V - V^*}{1 + \frac{d}{1}} p'(z) = 0. \quad (ii)
\]

The first condition simply says that price maximizes one period profits, given the level of \( z \). The second condition has two terms; the first, \( \pi_z \), is the loss in current profits resulting from increasing \( z \). The second is the present value of the difference in industry profits from next year on, multiplied by the marginal effect of \( z \) on the probability of entry. The second term is positive, and \( \pi_z < 0 \); the industry invests in \( z \) beyond the one-period optimum (that is the optimal level of \( z \) if entry is not threatened). Price is then set to maximize profits. The extent of the investment beyond the optimum is affected by several factors. One is the magnitude of the difference \( V - V^* \), the cost of having entry occur. The second is the discount rate. A third is the effectiveness of investment in \( z \) in reducing the probability of entry.

The effect of entry threats on price depends upon the sign of \( \pi_{qz} \). We know that the threat of entry increases \( z \). From condition (i)

\[
\frac{dq}{dz} = -\frac{\pi_{qz}}{\pi_qz}
\]

This derivative has the sign of \( \pi_{qz} \). Thus if \( \pi_{qz} > 0 \) entry will increase the price. This might well be the case if \( z \) were advertising, for advertising can decrease the demand elasticity and raise the profit maximizing price. On the other hand, if \( \pi_{qz} < 0 \), the threat of entry lowers price, as we observed earlier in the case of capacity.

Entry is deterred by investments that reduce the anticipated rate of return of the potential entrant and thus the probability of entry. These entry deterring investments proceed beyond the profit maximizing point, if entry is not a problem. They also affect pricing and output in a manner that depends upon the cross partial, \( \pi_{qz} \). If the latter is positive so that the investment increases the return to price increases, price rises. If increases in \( z \) lower the return to price increases as is true of the capacity case, price will fall.\(^{10}\)

A final remark about advertising may be in order. For a variety of reasons, industries have difficulty holding advertising levels down to those that maximize profits. The incentive structure of the game is that of the prisoner’s dilemma, and tacit collusion is difficult because of the presence of considerable exogenous randomness in the payoffs.\(^{11}\) However, the inability to collude at this level may contribute positively to a partial solution of the entry problem. For the deterrence of entry, as we have seen, requires investment above profit maximizing levels.

---

\(^{10}\) In the case of capacity \( z = k \) and \( \pi(q,k) = qD(q) - c(D(q),k) - rk \). Thus

\[
\pi_{qk} = -c_{zh}D'(q) < 0.
\]

\(^{11}\) Difficulties in tacit collusion with limited information and randomness are discussed in Spence (1974) and Stigler (1968).
existing industry on the other. The latter include more or less irreversible investments in a variety of kinds of capital. In a homogeneous product industry, a natural candidate is capacity, though that does not preclude other factors like a distribution system. In a differentiated product industry, advertising and other marketing activities that affect demand and raise the ante for the entrant also have the required effect. The irreversibility of the investment is important for two reasons. One is that it is a way for the existing industry to commit itself in advance, a way to issue a credible threat. Secondly, there is no need suboptimally to set a relatively flexible instrument like the price, since that can be adjusted within the time horizon required for entry to take place. Moreover, with differentiated products, it is not clear that prices could deter entry even if they were inflexible. Certainly as compared with the homogeneous case, the power of price to deter entry is attenuated.

Price might affect the probability of entry if it is taken as a signal by the potential entrant of the opportunities open to him. Therefore, the question is whether it is a good signal or not. High markups tend to be observed when demand elasticity is low. But with low elasticities of demand, increases in capacity and output would cause large price reductions. Thus, a high markup is not by itself a clear indication that entry would be profitable.

The implication of the view of entry taken here for welfare economics is the threat that entry does not necessarily improve resource allocation. The price may not fall. In some instances it may rise. If it does fall, it may not fall to the limit price. Production and distribution may be inefficient because of the overinvestment in the capital that deters entry. Exclusive dealerships seem a particularly common and striking example of this kind of problem.

References


