A Technological and Organisational Explanation for the Size Distribution of Firms*

by

Joshua S. Gans and John Quiggin**

First Draft: 22nd December, 1995
This Version: 11th August, 2000

This paper combines insights from the literature on the economics of organisation with traditional models of market structure to construct a theory of equilibrium firm size heterogeneity under the assumption of a homogenous product industry. It is possible that configurations consisting entirely of small firms (run by entrepreneurs with limited attention) and with larger firms (using managerial techniques to substitute away these limits to allow increasing returns technologies to become profitable) can arise in equilibrium. However, there also exist equilibrium configurations with the co-existence of large and small firms. The efficiency properties of these respective equilibria are discussed. Finally, the implications of an expanding market size are considered. *Journal of Economic Literature* Classification Numbers: D23 & L11.

Keywords: firm size, entrepreneur, increasing returns, strategic groups, bimodal distribution.

---

* We thank Susan Athey, Maxim Engers, Murray Kemp, Steven Klepper, Rodney Maddock, Rohan Pitchford, Scott Stern, Nancy Wallace, and participants at the Australian Economic Theory Workshop (La Trobe) for helpful discussions and comments on earlier versions of this paper. Responsibility for all errors lies with us.

** Melbourne Business School, University of Melbourne and Department of Economics, Australian National University, respectively. All correspondence to: Joshua Gans, Melbourne Business School, 200 Leicester Street, Carlton, Victoria 3053, Australia; Fax: (03) 9349-8169; E-mail: J.Gans@mbs.unimelb.edu.au.

It is a commonly observed characteristic of virtually every industry that firms of different sizes co-exist for long periods of time (see Simon and Bonini, 1958; Audretsch, 1995; and Sutton, 1997, 1998). This empirical phenomenon contrasts with traditional industrial organisational theory that predicts that firms with access to the same technologies will, in a competitive environment, operate at a similar scale of operation.[1] Such considerations form the basis of competitive market theory (Viner, 1931; Stigler, 1958) and of the more recent theory of contestable markets (Baumol, Panzar and Willig, 1982).

The competitive model does not provide an explanation of firm size heterogeneity because it is based on an implicit assumption that all technologies and organisational arrangements are infinitely reproducible. Under such an assumption, all firms use the technology generating the lowest average costs. However, when one technology or organisational arrangement uses an essential entrepreneurial input (or a factor supplied by an increasing cost industry), reproduction cannot be undertaken without rising costs. Thus, average costs involved in using that technology will rise as output expands. Only a few firms might be able to enter using that technology, leaving the market to other firms with technologies of different minimum efficient scales. In this paper, we show that allowing for limited reproducibility of this kind can generate equilibrium firm size heterogeneity in a competitive/free entry setting.

By using a competitive framework, our approach stands in contrast to previous theoretical work on the size distribution of firms. Ijiri and Simon (1977) tackled the regularity of industry configurations with a few large firms and many small firms by using stochastic methods in association with Gibrat's Law, that is, the observation that growth rates of firms were independent of their size. They showed that a purely stochastic

[1] This prediction is less sharp under constant returns to scale (Marris and Mueller, 1980).

mechanism[2] could account for many of the stylised facts regarding the observed size distribution of firms in many industries. Sutton (1998) has built on the stochastic approach. He views firms as clusters of investment opportunities such as plants, outlets and products that are at least as likely to be integrated in large as in small firms and incumbent firms as opposed to entrants. Under such conditions, he shows that the distribution of firm sizes will be skewed with industry assets concentrated in the hands of a few firms. Sutton then demonstrates that the empirically observed skewness is consistent with this viewpoint.

Not all of the features of the observed distribution of firms appear to be consistent with these stochastic models. They are based on the assumption that not only the mean but also the distribution of growth rates is unrelated to firm size. As a result, the distribution of firm sizes they generate is unimodal (typically lognormal). In many industries, such as retailing and restaurants, we observe a seemingly bimodal pattern with many one and two-establishment enterprises and a small number of large enterprises operating hundreds of establishments. In the service industries, large enterprises take the form of one or more chains of small establishments, operated either on a franchise basis or fully owned by the parent company. The stochastic approach would appear to predict larger numbers of middle-sized enterprises, say, those operating between two and one hundred establishments, than is in fact the case for those industries.

A similar objection can be raised with regard to the competitive models of Lucas (1978), and Calvo and Wellisz (1980). They postulated an underlying distribution of entrepreneurial or managerial talent in the economy and then showed that purposeful selection by agents into worker and managerial roles, combined with an assumption of Gibrat's Law, could provide a competitive foundation for heterogeneous firm sizes in

[2] In line with his arguments concerning the importance of satisficing, Simon regarded the absence of a competitive optimisation framework as a virtue.

equilibrium. This heterogeneity was, however, fully accounted for by the distribution of entrepreneurial ability. In the absence of detailed information about distributions of the ability of entrepreneurs attracted to different industries, their model had limited predictive power.[3]

In contrast to either the stochastic process or static equilibrium approaches, Jovanovic (1982) examines the firm size issue from a dynamic perspective. He argues that when firms begin production they discover potentially useful information regarding their costs; in particular, the costs associated with expansion. Those firms receiving favourable cost signals grow and become large firms. Those receiving unfavourable signals exit. At any point in time, therefore, large established firms and small firms who may only be there temporarily may populate an industry. This theoretical line can explain a potentially bimodal firm size distribution.[4]

In this paper, we combine insights regarding the technology of firms and the economics of organisation to build a simple but, we believe, very general theory of why one should *expect* heterogeneity in firm size and, in particular a bimodal size distribution, in many industries. In so doing, we assume that free entry and exit exist in a strong form, in that firms can be established using any available technology or organisational form. There are no barriers to production based on differential resources or exclusive access to key technologies – established firms and potential entrants are symmetric. In addition, our model is of an industry with a homogenous product, ruling out explanations for firm size differences based on niche exploitation or quality search considerations (see Salop and Stiglitz, 1977; and Bagwell and Ramey, 1994). Given our focus on the choice of low cost

[3] A similar result is contained in Kihlstrom and Laffont (1979). Calvo and Wellisz (1980) allow for some learning by agents before they become entrepreneurs but the size distribution of firms depends on a prior heterogeneity of talent in the population.

technology and organisational form by firms, assuming no start-up barriers and a homogenous product puts the maximum pressure on costly production methods to be excluded from production.

As a result, and in contrast to Jovanovic (1982), our theory is a static equilibrium approach. It can, therefore, explain the co-existence of large and small firms in an industry without requiring differences in survival or learning processes associated with entry. Our theory's value is in its simplicity and its potential to rationalise the persistence of firm size distributions across almost all industries; including those that may be more mature and where small firms may persist but stay small.

Section I outlines a basic model with these elements. There we postulate two organisational modes. First, there are *entrepreneurial* firms. These are managed by a single entrepreneur. Because of this restriction, any technology managed by such firms has a limited efficient scale.[5] This limits their size and also means that they will operate as price takers, entering production so long as revenue outweighs input costs and compensates the entrepreneur for the opportunity cost of their effort. We do assume, however, that entrepreneurial opportunities for a given industry are limited. That is, entrepreneurs can start production in an industry by seizing an opportunity. However, there is heterogeneity among entrepreneurial firms, with some opportunities being more attractive than others or alternatively, some entrepreneurs facing lower costs in identifying and exploiting opportunities than others. This makes the use of this technology not perfectly reproducible on an industry-wide scale. Hence, the supply curve from entrepreneurial firms is rising.

---

[4] This dynamic theory has other empirical implications including a low survival probability for small firms but a higher growth rate for those that do survive. Empirical work has confirmed these patterns for small businesses (Audretsch, 1995).

[5] For an early expression of this idea see Kaldor (1934).

But this is not the only possible organisational mode. If it were, the size distribution of firms would depend solely on any underlying heterogeneity in entrepreneurial talent in the population. Instead, we recognise the possibility of using production technologies exhibiting increasing returns to scale or subadditivity in costs (Murphy, Shleifer and Vishny, 1989; Romer, 1994). Such technologies often require large scales of production to become profitable and hence, management of such firms by an individual entrepreneur will not, in general, be optimal. An organisational form akin to Williamson's (1981) M-Form or *managerial firm* is required. That mode substitutes away from entrepreneurial inputs towards delegated managerial decision-making.[6] Firms using increasing returns technologies are unconstrained by limits on entrepreneurial abilities, but this freedom is gained at the cost of coordination and incentive difficulties within the organisation. These difficulties raise the average costs of such firms at all production scales. However, the managerial firm can adopt technologies that yield plant-level and enterprise level scale economies and can also replicate plants at cost-minimising scales of production. As such, long run average costs need not rise as the scale of output increases.[7]

Subject to the limited supply of entrepreneurial opportunities, we allow any firm to enter into production with any available technology and organisational mode. In our simple model, entrepreneurial firms are optimal for small scales of firm production while a managerial firm is optimal for large-scale operations. This is because of a complementarity between the entrepreneurial organisational mode and nonincreasing returns to scale (or backstop) technologies and between managerial firms and increasing returns technologies.[8]

[6] As described by Williamson (1981), Winter (1984) and Chandler (1990).

[7] There is a literature on how incentive and coordination problems can limit firm size. See, for instance, Arrow (1974) and McAfee and McMillan (1995). However, technological considerations can offset incentive and coordination problems. All that is necessary for the results below is that the maximum efficient scale of the largest firms in the industry is large relative to the market.

[8] This type of complementarity was emphasised by Winter (1984) in his distinction between 'entrepreneurial' and 'routinised' technological regimes.

Therefore, industry configurations consisting purely of entrepreneurial firms in perfect competition and a managerial firm that is a pure monopoly are possible equilibria. However, it is the interaction between firms operating with these two organisational modes and technologies that is of most interest here. Equilibria with a large number of entrepreneurial firms and a single managerial firm are possible. Moreover, the sustainable industry configuration could involve perfect but inefficient competition. This is despite the existence of a technology that would ordinarily result in a contestable natural monopoly.[9]

In many ways our basic story here is a formalisation of Porter's (1980) analysis of strategic groups within industries.[10] Like this paper, he notes that in many industries firm sizes are characterised by a bimodal distribution with a group of larger firms that have full product lines and high rates of vertical integration and a group of smaller firms that occupy niche areas and produce in low volumes. The reason smaller firms do not grow to be in larger firms arises because the large firms utilise technologies and other strategic variables that create entry barriers for firms of a similar type. However, opportunities do exist for small-scale entry, even despite the existence of entry barriers to large firms. Our model, however, abstracts from the complexity of Porter's approach by focusing on the homogenous products case and not explicitly considering issues of vertical integration allowing us to focus on the key assumptions generating separate strategic groups within an industry.

The remainder of the paper proceeds as follows. Section III looks at the efficiency of sustainable configurations while, in Section IV, we conduct comparative statics on

[9] The model presented here is related to dominant firm models in industrial organisation (Tirole, 1988). Such models involve a dominant firm interacting with a competitive fringe that exists because of its control over natural resources or ability to enter with positive capacity. The difference here will be the critical role of entrepreneurs. This is important because the model presented in this paper accounts for persistent firm size heterogeneity. Control over natural resources and capacity are imperfect sources of such persistence especially given the possibility of merger. As will be shown in Section III, there are complications to such procedures where entrepreneurial resources are concerned.

sustainable industry configurations as the size of the market increases and show that industries will evolve from the purely entrepreneurial to mixed and then eventually, a pure monopoly. A final section concludes.

I Firm Size and Organisational Mode

We begin by presenting and motivating a set of assumptions regarding the interactions between firm size, technological choice and organisational mode. The purpose here is to explore the nature of cost functions in alternative organisational modes that will be used in our general theory of equilibrium firm size heterogeneity. We will show that the organisational mode for an individual firm changes with its desired size. In the next section, we go on to endogenise firm size within an industry equilibrium model.

There are two types of firms. Firms that are owned and managed by entrepreneurs are termed *entrepreneurial*. These firms use entrepreneurial effort to enhance productivity. Entrepreneurs face a strictly convex opportunity cost of effort, $g_i(e_i)$, where e_i is the total effort supplied by entrepreneur i. Marginal effort cost is assumed to approach infinity as effort approaches some finite maximum.[11] However, it is assumed that the conditions under which entrepreneurial effort is supplied are such that each firm will only employ a single entrepreneur. This means that the index i identifies the entrepreneur as well as the firm. We do not model the rationale for this assumption here,[12] simply capturing it by assuming the

[10] See also Caves and Porter (1977).

[11] This is a reasonable assumption given that entrepreneurs will have only a finite amount of time to devote to managerial tasks such as monitoring (Gifford, 1992).

[12] This assumption can be justified by conceiving the entrepreneur as supplying an input that cannot be monitored or at least cannot be perfectly monitored. For a firm with a single entrepreneur this difficulty is overcome by making the entrepreneur the recipient of residual income. Incentive problems arise when multiple entrepreneurs form a partnership (Hart and Moore, 1990; Abreu, Milgrom and Pearce, 1991). Further problems arise if partners differ in ability but an equal sharing rule is imposed (Farrell and Scotchmer, 1988).

entrepreneurial effort costs are superadditive. All of these assumptions are summarised as (A1):

(A1) $g_i(e_i)$ is strictly quasi-convex and continuously differentiable with $\lim_{e_i \to \bar{e}_i} g_i'(e_i) = \infty$. Total costs from employing more than one entrepreneur are superadditive, that is, $g_i(e) < g_i(e_i) + g_j(e - e_i), \forall e, e_i, i \neq j$.

The second type of firm we term *managerial*. Such firms do not find it optimal to employ entrepreneurial effort in production. The existence of incentive contracts or efficiency wage payments that mitigate the need for monitoring mean the entrepreneurial input is not necessary for labour management purposes.[13] Therefore, managerial firms are not constrained by the limited attention of the entrepreneur.

There are two production technologies available to all firms. These technologies produce goods that are perfect substitutes. The entrepreneurial technology,[14] subscripted by E, is a *backstop* technology that exhibits constant returns to scale and is reliant on entrepreneurial effort: $q_i = \min[q_E l_i, e_i]$, where l_i is the amount of labour devoted to this production process, q_E is the marginal productivity of labour using the entrepreneurial technology, q_i is the output of firm i, and e_i is the level of effort supplied by the entrepreneur. Thus, total cost for firm i using the entrepreneurial technology is given by, $C_E(q_i) = \min_{e_i} \{w_E q_i / q_E + g_i(e_i)\}$ where w_E is the wage paid by such firms. Under (A1), average cost, $AC_E(q_i)$, is either everywhere increasing or U-shaped.

The managerial production technology, subscripted by M, does not use entrepreneurial effort as an input. Instead, it is an *increasing returns* to scale technology using only labour: $q_i = q_M(l_i - F)$ where F represents a fixed input requirement. For simplicity, as in many models in industrial organisation (Krugman, 1990), we assume that this

---

[13] This is demonstrated in the Appendix extending a model from Milgrom and Roberts (1992, Chapter 8).
[14] So named to anticipate the fact that entrepreneurial firms will be the only firms using this technology.

takes the form of pure labour overhead. Costs for this technology are given by $C_M(q_i) = w_M(F + q_i / q_M)$.

With a constant wage, average costs for managerial firms are decreasing everywhere (i.e., display subadditivity). However, marginal labour costs, $w_M(q)$ and $w_E(q)$, both potentially depend on the level of output. The reason for this could be span of control issues (McAfee and McMillan, 1995; Qian, 1994; Aghion and Tirole, 1995), coordination issues (Kaldor, 1934; Penrose, 1959; Arrow, 1974; Becker and Murphy, 1992), influence costs (Meyer, Milgrom and Roberts, 1992) or monitoring costs (Williamson, 1981). These could constrain the size of managerial firms f their rate of increase was faster than the technological gains from scale economies. However, both the ability to adopt holding company models and the possibility of substituting incentive or efficiency wage contracts for monitoring mean that unit labour costs are potentially bounded from above. In the appendix, we analyse this possibility in a simple efficiency wage model. Although it is rationalised by that model, here, we simply make the assumption that wages in the managerial firm rises at a sufficiently low rate to generate declining long run average costs.

(A2) $$\frac{w_M(q_i)}{w'_M(q_i)q_i} > \left(F + \frac{q_i}{q_M}\right)\frac{1}{F}.$$

Figure 1: Entrepreneurial Versus Managerial Average Costs

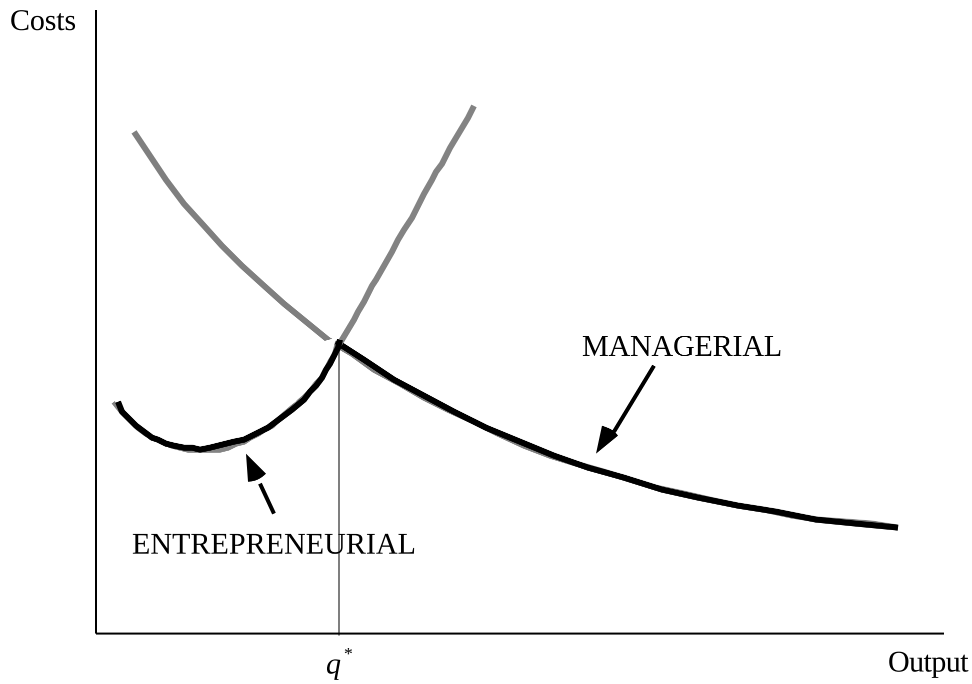


Figure 1 shows the average cost curves for the two technologies. Up to some minimum output, q^*, average cost is lower for the entrepreneurial technology, but cost-minimisation beyond this point requires adoption of the managerial technology. In this figure, the entrepreneurial technology actually has higher minimum average costs than the managerial technology for high output levels. This need not be the case. If the opportunity costs of entrepreneurial effort are low, then average costs for an entrepreneurial firm might be lower than that of a managerial firm operating at a high output level. Alternatively, this could be the case if the entrepreneurial technology is more efficient at the margin.

II Industry Equilibrium

The previous section stated assumptions under which there is a non-convexity in average costs for a given firm. Firm output is determined endogenously as part of an industry equilibrium. In this section, we analyse equilibrium outcomes using the notion of a sustainable industry configuration as defined by Baumol et.al. (1982).[15]

We consider an industry that produces a homogeneous good, the market demand for which is described by a continuous function $D(\{p_i\}_{i\in I};f)$ where p_i is the price of firm i, I is the set of firms in the industry and f is a parameter that describes the size of the market. We assume:

(A3) D is nonincreasing in each price and is increasing in f.

The Entrepreneurial Fringe

We assume that entrepreneurial opportunities are a scarce resource for the industry:[16]

(A4) g_i is nondecreasing in i.

Because they use a constant returns to scale technology, we model entrepreneurial firms as competitive, taking the price, p, of the good as given. By (A1) and (A4), as the number of firms in the industry rises, the marginal cost of effort for the marginal firm in the industry is nondecreasing.

The entrepreneurial fringe is defined as the set of all firms adopting the entrepreneurial technology, whether or not their output is positive at any given price. (A4) implies that the long run supply curve for the entrepreneurial fringe of the industry is rising. Formally, we can express the total quantity supplied by the entrepreneurial fringe in the long-run as a nondecreasing function, $S(p)$ of the market price, p.[17] The qualitative properties of the long run competitive supply curve play a critical role in the results that follow.

[15] This approach was developed earlier in the international trade literature, see Kemp (1969, p.155).
[16] Alternatively, the entrepreneurial technology is not replicable without efficiency losses.
[17] In graphical illustrations below, $S(p)$ is a continuous function. The theoretical analysis below does not require such an assumption.

The assumption that the long-run supply curve for the entrepreneurial fringe, while critical to our argument, is a relatively weak assumption. It can be supported by many alternative conditions similar in effect to (A4).[18] Basically, we are ruling out situations where there are unlimited opportunities for entrepreneurs to enter the market or even where those opportunities are unlimited, different entrepreneurs have different abilities and capabilities. Hence, the more productive entrepreneurs will enter the industry first.

Managerial Firms

Managerial firms use the increasing returns to scale technology. Therefore, assuming (A2), they exhibit falling long run average costs. Hence, if this were the only prevailing technology, the industry would be traditionally characterised as a natural monopoly.[19] Only one firm with access to this technology will enter into production in the industry. For convenience, we will refer to this managerial firm as firm 0. This firm could be a price maker. However, for the moment, it will be assumed that the constraint of potential competition ensures that firm 0 will set price equal to average cost. Therefore, any industry configuration will have at most one managerial firm and if this firm produces positive output,

$$p = \frac{w_M l_0}{q_0} = w_M \left(\frac{F}{q_0} + \frac{1}{q_M} \right).$$

Since the good produced by both types of firm is homogenous, the residual demand faced by this individual firm can be represented by, $d(p;S) = D(.) - S(p)$.

---

[18] For instance, Kihlstrom and Laffont (1979), in a model based on Knight (1921), assume that potential entrepreneurs differ in terms of their degree of absolute risk aversion. Since entrepreneurs are residual claimants on firm profits in excess of labour costs, at low price levels, only the least risk averse entrepreneurs set up firms. Lucas (1978) and Gifford (1993) provide alternative rationales supporting this condition.

[19] See Baumol et. al. (1982) and Tirole (1988).

Sustainable Industry Configurations

An industry configuration is a list of prices and outputs for each firm *i* in the industry.[20] Here we list the standard definition of a sustainable industry configuration and a sustainable natural monopoly.

Definition (Sustainable Industry Configuration). *An industry configuration,* $\{(\hat{p}_i, \hat{q}_i)\}_{i \in I}$*, is sustainable if,*

(i) industry supply equals market demand, Q = D(.);

(ii) all active firms earn nonnegative profits, i.e., $p_i(p_i, q_i) \geq 0$, $\forall i$ *with* $q_i > 0$*;*

(iii) all inactive firms (taking prevailing prices as given) earn nonpositive profits on entry.

Definition (Sustainable Natural Monopoly). *A sustainable industry configuration is a natural monopoly if* $p_0(p_0, q_0) \geq 0$ *and* $p_i(p_i, q_i) < 0$, $\forall i \neq 0$.

It may be that no sustainable industry configuration exists for a given market demand, $D(.)$. Baumol et.al. (1982) provide examples of non-existence based on the integer problem. However, given the residual demand curve facing the managerial firm, the existence of a sustainable allocation in our problem is equivalent to the existence of a sustainable allocation in the problems examined by Baumol et.al. (1982). In particular, in the single-good case, if average cost is everywhere decreasing, a sustainable allocation with at most one managerial firm must exist. Note that whereas in Baumol et al (1982), sustainable allocations with zero managerial firms are trivial, in our model they correspond to competitive allocations.

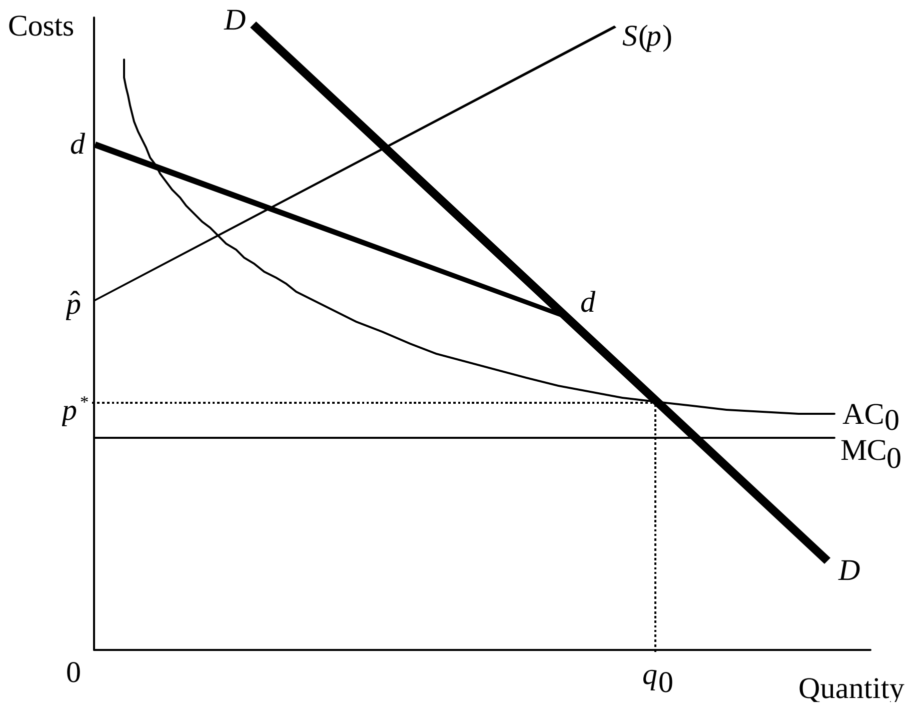
With this we can state the following necessary and sufficient conditions for the industry to be a sustainable natural monopoly.

Proposition 1. *Assume (A1) - (A4). Let* $\hat{p} \equiv \arg\max\{p | S(p) = 0\}$ *and let* $p^* \equiv \{p | (p - w_M q_M^{-1}) d(p;S) = w_M F\}$*. Then, if a sustainable industry configuration exists, it is a sustainable natural monopoly if and only if* $p^* \leq \hat{p}$.

[20] We use the language 'industry configuration' rather than industry or market structure, following Baumol et.al. (1982).

Under this condition, the price at which the managerial firm breaks even is lower than the highest price that would allow positive output by the entrepreneurial fringe (Figure 2).[21]

Figure 2: A Sustainable Natural Monopoly



In addition, we can state necessary and sufficient conditions for the sustainable industry configuration to be composed solely of entrepreneurial firms (Figure 3).

Proposition 2. *Assume (A1) - (A4). Let* $d^{-1}(q_0;S)$ *be the inverse demand function for firm 0. Then a necessary and sufficient condition for* $q_0 = 0$ *in any sustainable industry configuration is that* $d^{-1}(q_0;S) < w_M\left(\frac{F}{q_0} + \frac{1}{q_M}\right)$, *for all* q_0.

Figure 3: A Sustainable Configuration with Entrepreneurial Firms Only

[21] In Figures 2 through 6, we have reversed the axes so that the graphs are closer to the familiar textbook representations. Nonetheless, for ease of reference we continue to label the fringe supply, market and residual demand curves as $S(p)$, $D(p)$ and $d(p)$ respectively.

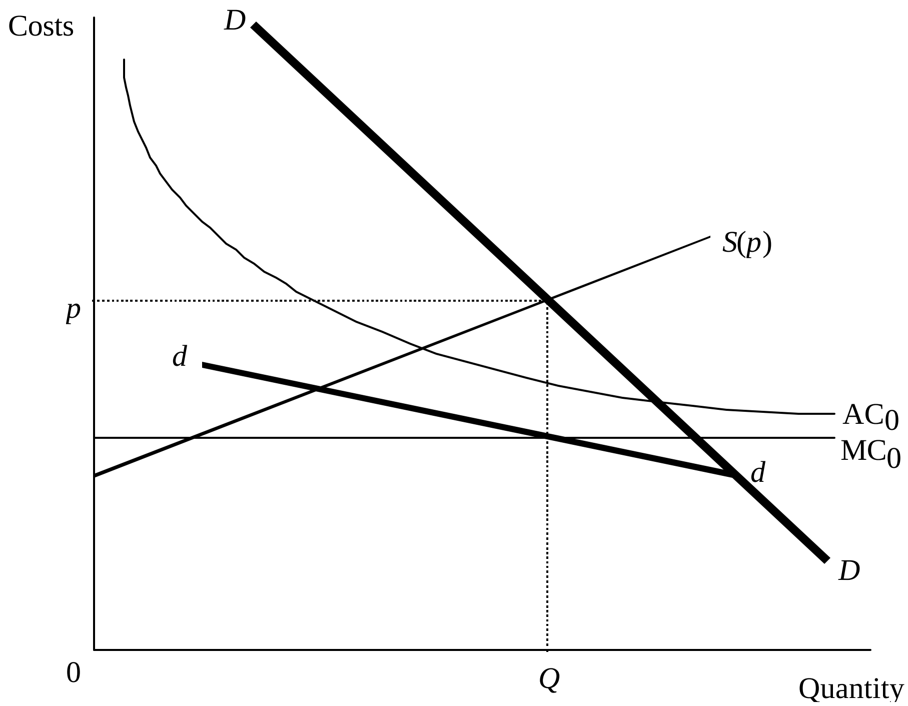
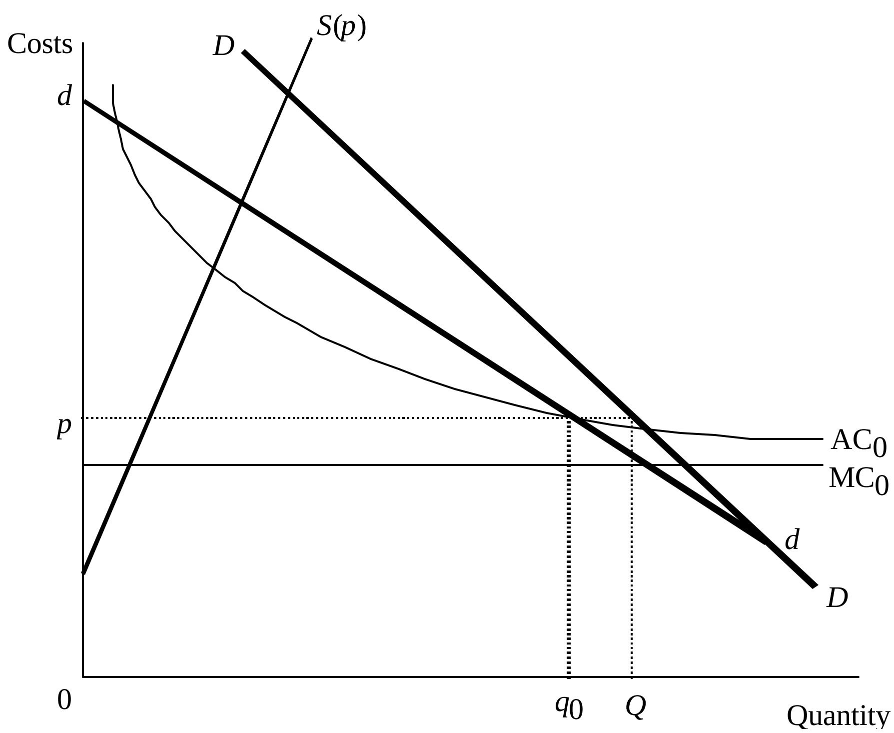


Figure 4: A Mixed Sustainable Configuration



This condition means that residual demand lies below the long run average cost curve of the managerial firm, not allowing it to break even at any output level (Figure 3). This will occur if wages and fixed costs are high, and the marginal product of labour is low relative to entrepreneurial firms.

The conditions on market demand and fringe supply curves for the pure cases of Propositions 1 and 2 are potentially disjoint. Provided that the fringe supply curve is of a particular form, mixed configurations, comprising a managerial firm and a number of entrepreneurial firms are, therefore, possible.

Proposition 3. *Assume (A1) - (A3). A necessary condition for there to exist a market demand, D(.), that supports a mixed industry configuration with S(p) > 0 and* $q_0 > 0$*, is that S(p) be strictly increasing for some p.*

An increasing fringe supply generates the possibility of sustainable industry configurations with both entrepreneurial firms and a managerial firm (see Figure 4). If entrepreneurial opportunities are unlimited (that is, (A3) did not hold or $\partial^2 g_i / \partial i = 0$), then the long-run supply curve from entrepreneurial firms is flat, with all entrepreneurial firms producing at minimum average cost. In this case, the only possible industry configurations are those composed of entirely of entrepreneurial firms or a natural monopoly. On the other hand, mixed configurations become more likely when there is a limited supply of quality entrepreneurs. In this case, while fringe supply might be positive for low prices, it rises steeply as entrepreneurs with low opportunity costs become scarce.

Other Large and Medium Size Firms

The above model makes an explicit prediction that if a mixed configuration occurs it consists of a single large firm and many smaller firms. All of the results above hold for more general models including those in which the managerial firm is a multi-product firm and there is more than a single input into production. The key conditions are technological -- that entrepreneurial firms have U-shaped average cost curves with low minimum efficient scales relative to the size of the market and that the managerial firm has access to an increasing returns technology so that its minimum efficient scale is large relative to the size of the market. The former condition ensures competition among entrepreneurial firms and the latter ensures that only one managerial firm enters the market. All of the above results would apply, without qualification, to general technologies of that kind.

To see this, suppose that there exists “intermediate” technologies in which firms can choose their level of fixed costs, with greater levels of fixed costs enhancing labour productivity. That is, q_M is a function of F with $q_M(F') \geq q_M(F)$ for $F' > F$.[22] Despite the endogeneity of F, these technologies display subadditivity in costs and hence, under our assumptions (effectively, of Bertrand competition), in any sustainable industry configuration, only one firm will enter incurring the level of fixed costs that minimises average costs subject to output generated by the entrepreneurial fringe. Hence, to amend the model to allow for medium size firms and other large firms we need to look elsewhere.

By relaxing the constant returns assumption for entrepreneurial firms, it is possible to consider monopolistic competition among firms in the entrepreneurial fringe and, by providing limits on the extent of scale economies, we can allow for natural duopoly or oligopoly in the managerial firm sector. Alternatively, the assumption of Bertrand equilibrium implicit in the use of the sustainability concept could be replaced by an alternative such as Cournot equilibrium, which could support two or more firms, assuming that anti-trust or strategic considerations precluded merger. The intricacies of such cases are, however, beyond the scope of the present work. They would require a consideration of the public good and strategic considerations in forestalling entry and in merger decisions (see Salant, Switzer and Reynolds, 1983).

Of critical importance in many of the above results is the elasticity of the fringe supply curve, $S(p)$. Many textbooks describe cases in which this function is perfectly elastic (see Mas-Colell, Whinston and Green, 1995, pp.334-339). Under such an assumption, there are only two types of industry configurations -- entrepreneurial competitive equilibria and sustainable natural monopolies (see Proposition 3). However, there are several alternative ways in which $S(p)$ could be upward sloping in addition to the entrepreneurial differences assumed here. Suppose that managerial firms produce a slightly better quality good or a good that appeals to a large number of consumers. Then the managerial firm has some monopoly power over the good. This could be the result of sunk costs in advertising (Sutton, 1991), product innovation (Sutton, 1996), or product differentiation (von Ungern-Stenberg, 1988). In this case, the monopolist supplies the differentiated product at a different price from the fringe's product but both supply positive amounts. The fringe's supply curve is upward sloping because competition among entrepreneurial firms involves close but imperfect substitutes. But their products are less substitutable for the managerial firm's product. Stern (1995) and Bresnahan, Stern and Trajtenberg (1997) provide empirical evidence for low degrees of substitutability among goods supplied by large as opposed to fringe firms (that is, branded versus generic suppliers). Thus, the incorporation of product differentiation would enhance the conclusions of this paper. Nonetheless, the key driving force for heterogeneity among firms remains technological and organisational rather than preference related.

III Efficient Industry Configurations

As a response to the conceptual and practical difficulties in establishing the optimal size of firms in an industry, Stigler (1958) formulated the "survivor principle." That principle asserted that, if firms of a given size were observed to continue production in an industry

---

[22] Dasgupta and Stiglitz (1980) employ this technological set-up.

over a period of time, this would be considered the optimal firm size. Stigler quotes John Stuart Mill as an antecedent to this principle:

> Whether or not the advantages obtained by operating on a large-scale preponderate in any particular case over the more watchful attention, and greater regard to minor gains and losses usually found in small establishments, can be ascertained, in a state of free competition, by an unfailing test.... Whereever there are large and small establishments in the same business, that one of the two which in existing circumstances carries on the production at the greater advantage will be able to undersell the other. (Mill quoted by Stigler, 1958, p.27).

Stigler then argues that oligopolistic industries too, if contestable, will eliminate firms of inefficient size.

The model of this paper, by showing how a distribution of firm sizes is determined endogenously, can shed light on the applicability of the survivor principle and other theories that posit that free entry and exit leads to least cost production in homogenous product industries (e.g., Baumol et.al., 1982). In this section, we analyse whether observed industry configurations are, in fact, efficient.

There is, potentially, a difference between efficient and sustainable industry configurations. First, we need to state what we mean by efficiency.

Definition (Efficiency). *An efficient industry configuration maximises consumer surplus in the industry subject to the requirement that firms break even and charge linear prices.*

This definition neglects any entrepreneurial rents.[23] It is also different from an unconstrained definition of efficiency without break-even or linear price constraints.[24] Efficiency without such constraints would involve all units being produced with the minimum marginal cost. Thus, it could involve some "good" entrepreneurial firms producing a fraction of demand

---

[23] Suppose that the losses in consumer surplus arising in any sustainable industry configuration were greater than the value of entrepreneurial rents earned. If the managerial firm could "buy out" the fringe and prevent their entry, would efficiency be restored? The possibility that any potential managerial firm means that entrepreneurial firms will receive all of the surplus generated from this buy out. As such, even if the resulting monopoly was constrained to limit price at the original equilibrium price, the total quantity sold in the market would be the same allowing for the buy out. So an efficient solution would not result as the amount of the buy out raises the managerial firms fixed costs. The precise form that a buy out would take is an area open for future research.

and the managerial firm supplying the rest. Typically, however, this equilibrium is not sustainable in the absence of some subsidisation of the managerial firm.

For any given market demand, a natural monopoly is an efficient industry configuration in the sense above if $w_M\left(\frac{F}{D(.)}+q_M^{-1}\right)\leq S^{-1}(D(.))$, that is, if the managerial firm can supply the entire market demand at a lower price than the entrepreneurial fringe and still break even. This observation leads to the following proposition.

Proposition 4. *Assume (A1) - (A3). If there exists an industry configuration that is a sustainable natural monopoly then that configuration is efficient.*

> PROOF: Suppose that $w_M\left(\frac{F}{D(.)}+q_M^{-1}\right)>S^{-1}(D(.))$; i.e., a natural monopoly is not an efficient industry configuration. Suppose now there exists a $D(.)$ such that there is a sustainable natural monopoly. By proposition 1, for that $D(.)$, $p^*\leq\hat{p}$. Since $p^*=w_M\left(\frac{F}{D(.)}+q_M^{-1}\right)$ and $\hat{p}=S^{-1}(0)$, this implies that $w_M\left(\frac{F}{D(.)}+q_M^{-1}\right)\leq S^{-1}(0)\leq S^{-1}(D(.))$, with the only equality at $D(.)=0$. Therefore, for positive market demand there is a contradiction.

Note that the converse is not true (Figure 5). That is, it is possible for purely entrepreneurial configurations to be efficient. Therefore, fringe competition could drive out an efficient natural monopoly. This is the case in Figure 3.[25] Observe, however, that as $S(p)$ becomes more elastic, the possibility of an inefficient entrepreneurial configuration is diminished. Indeed, for a mixed configuration, the efficiency considerations are unambiguous.[26]

Proposition 5. *Assume (A1) - (A4). Any sustainable mixed configuration is inefficient.*

---

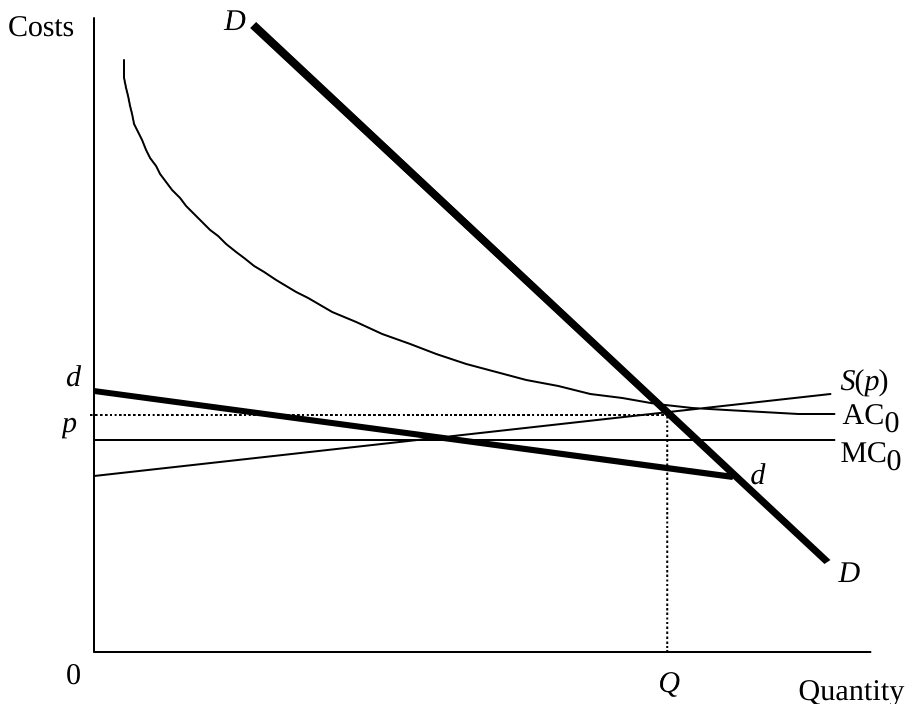
[24] This is a standard efficiency measure where firms are constrained to offer a simple linear price. Laffont, Rey and Tirole (1998) refer to this as a Ramsey pricing outcome.

[25] Our results do not preclude the possibility that an unconstrained monopoly could be more efficient than the free entry case.

[26] This result stands in contrast to Jovanovic (1982) who found that equilibrium firm size distributions were efficient.

This proposition follows directly from the assumption of falling average costs for the managerial firm and nondecreasing fringe supply. It implies that consumer surplus will be raised if entrepreneurial firms are restricted from entry into the industry.

Figure 5: A Sustainable and Efficient Equilibrium with Entrepreneurial Firms



In conclusion, this section shows that while observed monopoly configurations can be efficient it is possible that these might be subject to inefficient competition from an entrepreneurial fringe. Moreover, observed mixed configurations are always inefficient with a regulated single managerial firm providing greater consumer welfare. Note these results are under conditions favourable to contestable outcomes and are obtained without any strategic behaviour being undertaken. Thus, Stigler's (1958) observations of persistence of firms of a wide range of sizes does not necessarily support his conclusions that average cost curves are flat over a large range of output.

IV Evolution of Market Structure

Chandler (1990) has described in meticulous detail how many industries evolved from configurations consisting of many entrepreneurial firms to ones in which one or a couple of large managerial firms dominated market share in the industry. A natural question that arises in the context of the model developed in this paper is what factors might account for such changes? Here, we identify market size as a possible explanation of the observed evolution of market structure showing how the nature of sustainable industry configuration changes as the market grows.

Despite the fact that a multiplicity of sustainable industry configurations is possible, we can conduct general comparative statics analysis on industry configurations. We examine changes in the set of equilibria, in particular, the level of managerial firm output, q_0, as the size of the market increases.

Proposition 6. *Assume (A1) - (A3). Suppose that S(p) is nondecreasing. Then the set of* q_0 *characterising any sustainable industry configuration is nondecreasing in the market size parameter,* f.

> PROOF:[27] First observe that the left hand side of the condition for an industry condition involving $q_0 = 0$, $d^{-1}(q_0; S, f)$, is nondecreasing in f, while the right hand side is unchanging. Therefore, as market size increases we move from sustainable configurations with $q_0 = 0$ to ones where this is positive. Since $S(p)$ is determined purely by technological and entrepreneurial supply considerations, as we raise market size in any mixed configuration, the individual demand curve facing the managerial firm rises and given the declining average costs, so does q_0. This also holds for situations in which the industry configuration is a sustainable natural monopoly. Finally, note that as market size increases p^* falls while $\hat{p}$ stays unchanged.

[27] This is a sketch of the formal proof which requires a precise definition of movements in the set of equilibria as multiplicity is not ruled out by the above assumptions. This could be done by conducting comparative statics on the highest and lowest equilibria or on some more general notion of set ordering (see Athey, Milgrom and Roberts, 1996, for more details).

Thus, as the size of the market grows the industry becomes more concentrated with a dominant firm emerging. This evolution of sustainable industry configurations is summarised in Figure 6. The total output of entrepreneurial firms falls as the market grows. This is because a larger market raises the demand curve for the managerial firm and allows it to reduce price without making a loss. This reduction in price causes some entrepreneurial firms to leave the market. Thus, both the market share and output level of the entrepreneurial fringe falls.

Figure 6: Efficient and Sustainable Configurations for Different Market Sizes

V Conclusions and Future Directions

The size distribution of firms has received relatively little attention in the industrial organisation literature. In popular discussions of the economy a clear distinction is often drawn between ‘big business’ and ‘small business’. However, models of the size distribution of firms based either on stochastic growth processes or heterogeneity in entrepreneurial opportunities suggest that no distinction of this kind is justified. The properties of the Pareto distribution that arises from the stochastic Gibrat’s Law approach are such as to suggest that any point at which a division might be drawn is as good or as bad as any other point.

The model advanced here is based on a qualitative distinction between organisational modes rather than on continuous random variables. The key result is that, under a range of demand and cost conditions, multiple organisational modes can co-exist, creating a size distribution of firms quite different from that arising from random variation in growth rates or entrepreneurial ability.

The analysis here could be extended in a number of ways. The most interesting avenues involve introducing dynamics. Two dynamic processes that might be considered are the conversion of one or more entrepreneurial firms in an initially competitive industry into managerial enterprises and the diversification of existing managerial enterprises into industries previously populated by entrepreneurial firms.

A further avenue would be the introduction of a stochastic element, permitting the derivation of statistically testable predictions about the relationship between size and growth rates and about the equilibrium population distribution. We conjecture that firms in a medium-size output range, lying between the largest optimal output for an entrepreneurial firm and the minimum efficient scale for a managerial firm will display highly variable and highly skewed growth rates, pursuing in effect a strategy of 'get big or get out.'

Finally, although the model presents a solution to the puzzle of large and small firms, it does not appear to generate sustainable industry configurations involving more than one large firm coexisting with a fringe of small firms. As noted earlier, such configurations may perhaps be explained by considerations of product differentiation. Alternatively, it may be that observed configurations of this kind are not sustainable equilibria in the sense defined above. Profitable mergers between large firms may be prohibited under antitrust laws or discouraged by strategic considerations such as those examined by Salant, Switzer and Reynolds (1983).

Finally, the model presented here also has potential policy implications. The case of inefficient mixed equilibria may be seen as a link between the concept of unsustainable natural monopoly and the older debate over the potential for ‘cream-skimming’ in natural monopoly industries with open entry.

Appendix:

Here we consider an example where the organisational mode is modeled explicitly using the efficiency wage model of Becker and Stigler (1974), as amended by Milgrom and Roberts (1992). In this example, complementarities between technological and organisational choices give rise to the basic nonconvexity illustrated in Figure 1.

In addition to providing a direct input into production, the entrepreneur contributes to labour management. That is, the entrepreneur can devote effort to reducing total labour costs. To take a simple case, suppose that workers can potentially be employed but supply zero labour productivity, with q = 0, due to the possibility of "shirking." If workers do not shirk, q = 1. This underlying labour productivity function is common across entrepreneurial and managerial firms. Shirking workers, however, could be caught with probability r, in which case they lose their job and receive a reservation wage, $\overline{w}$. Suppose that the worker's private gain to shirking is g. Then there are two instruments by which entrepreneurs can deter shirking. First, they can employ workers for an *efficiency wage* that gives workers a premium for employment within the firm making them indifferent between shirking or not. That minimum efficiency wage is given by the following expression: $g \geq r(w - \overline{w}) \Rightarrow w^* = g/r + \overline{w}$. Second, they can raise the detection probability by devoting effort to *monitoring* the worker: $e_i^M = M(r)$, where e_i^M is the effort required to achieve a detection probability of r and M is a increasing, convex function with $0 = M(\underline{r})$. The latter assumption simply says that when no effort is devoted to monitoring there exists a lower bound on the detection probability of $\underline{r} > 0$. This means that monitoring is not necessary to ensure finite labour costs.

Given this, total labour costs are $(w + e_i^M)l_i$. Substituting in the efficiency wage function, these costs become solely a function of the choice of r: $(g/r + \overline{w} + M(r))l_i$. In terms of labour costs, a greater labour force and greater monitoring will be substitutes if $M_r(\underline{r}) \geq g/\underline{r}^2$. In this case, the unit cost of employing labour is everywhere (weakly) increasing in monitoring intensity. This is because unit monitoring costs are assumed to rise more than the reduction in the efficiency wage.

Given the technological specifications, when we consider the costs of entrepreneurial effort as well as labour costs, greater employment and greater monitoring intensity are substitutes almost everywhere. Consider first the case where the entrepreneurial technology is used. In this case, raising employment is complementary with greater effort devoted to production. Because effort costs are convex, effort in production is a substitute with monitoring effort. Therefore, employment and monitoring intensity are substitutes in entrepreneurial as well as labour costs and, as such, raising l_i reduces the optimal choice of r.

The same is true if the increasing returns technology is used. While the entrepreneurial effort substitution effect is no longer relevant here, employment and monitoring intensity are substitutes in terms of labour costs. Hence, larger employment will result in lower monitoring intensity. For a firm that is sufficiently large, no monitoring may take place at all if monitoring intensity is the same for all workers. If monitoring intensity is variable, monitoring effort is likely to become focused on a smaller proportion of workers. Either way, total effort devoted to monitoring will fall as the firm grows large.

Finally, if the technology switches from the entrepreneurial to managerial when employment increases, it is possible that monitoring intensity could rise during the transition. This is because the switch in technology relieves the entrepreneur of a direct production role and switches this to labour cost reduction. Locally, around the switching point of employment, monitoring intensity will rise and efficiency wages will fall. Note that this means that for a small range in employment, the opportunity cost of monitoring falls rather than rises. Thus, a nonconcavity in the returns to monitoring is introduced at this point.

As firm output grows, monitoring intensity falls. Therefore, the organisational mode switches from one that relies on an entrepreneurial role in production to a managerial mode that is not constrained by the limited attention of the entrepreneur. The possibility of switching to the increasing returns to scale technology and to switching to labour discipline devices, other than direct monitoring, means that firm size is unconstrained. A typical firm has a nonconvex average cost function as depicted in Figure 1.

References

Abreu, D., P. Milgrom and D. Pearce (1991), "Information and Timing in Repeated Partnerships," *Econometrica*, 59 (6), 1713-1734.

Aghion, P. and J. Tirole (1995), "Some Implications of Growth for Organizational Form and Ownership Structure," *European Economic Review*, 39, 440-455.

Arrow, K.J. (1974), *The Limits of Organization*, Norton: New York.

Athey, S., P. Milgrom and J. Roberts (1996), *Monotone Methods for Comparative Static Analysis*, (forthcoming).

Audretsch, D.B. (1995), *Innovation and Industry Evolution*, MIT Press: Cambridge (MA).

Bagwell, K. and G. Ramey (1994), "Coordination Economies, Advertising, and Search Behaviour in Retail Markets," *American Economic Review*, 84 (3), 498-517.

Baumol, W.J., J. Panzar, R. Willig (1982), *Contestable Markets and the Theory of Industry Structure*, Harcourt Brace Jovanovich: New York.

Becker, G. and K.M. Murphy (1992), "The Division of Labor, Coordination Costs and Knowledge," *Quarterly Journal of Economics*, 108 (4), 1137-1160.

Becker, G. and G. Stigler (1974), "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, 3 (1), 1-18.

Bresnahan, T., S. Stern and M. Trajtenberg (1997), "Market Segmentation and the Sources of Rents from Innovation: Personal Computers in the late 1980's," *Rand Journal of Economics*, 28 (0), pp.S17-S44.

Calvo, G.A. and S. Wellisz (1980), "Technology, Entrepreneurs and Firm Size," *Quarterly Journal of Economics*, 95, 663-677.

Caves, R. and M.E. Porter (1977), "From Entry to Mobility Barriers," *Quarterly Journal of Economics*, 91 (2), 241-261.

Chandler, A.D. (1990), *Scale and Scope: The Dynamics of Industrial Capitalism*, Belknap Press: Cambridge (MA).

Dasgupta, P. and J.E. Stiglitz (1980), "Industrial Structure and the Nature of Innovative Activity," *Economic Journal*, 90, 266-293.

Farrell, J. and Scotchmer (1988), "Partnerships," *Quarterly Journal of Economics*, 103 (2), 279-297.

Gifford, S. (1992), "Innovation, Firm Size and Growth in a Centralized Organization," *Rand Journal of Economics*, 23 (2), 284-298.

Gifford, S. (1993), "Heterogeneous Ability, Career Choice and Firm Size," *Small Business Economics*, 5(4), December 1993, 249-259.

Hart, O. and J. Moore (1990), "Property Rights and the Nature of the Firm," *Journal of Political Economy*, 98 (6), 1119-1158.

Ijiri, Y. and H.A. Simon (1977), *Skew Distributions and the Sizes of Business Firms*, North-Holland: Amsterdam.

Jovanovic, B. (1982), "Selection and the Evolution of Industry," *Econometrica*, 50 (2), 649-670.

Jovanovic, B. and G.M. MacDonald (1994), "The Life Cycle of a Competitive Industry," *Journal of Political Economy*, 102 (2), 322-347.

Kaldor, N. (1934), "The Equilibrium of the Firm," *Economic Journal*, 154, 70-81.

Kemp, M.C. (1969), *The Pure Theory of International Trade and Investment*, Prentice-Hall: Englewood Cliffs.

Kihlstrom, R.E. and J-J. Laffont (1979), "A General Equilibrium Entrepreneurial Theory of Firm Formation Based on Risk Aversion," *Journal of Political Economy*, 87, 719-748.

Klepper, S. (1996), "Entry, Exit, Growth and Innovation Over the Product Life Cycle," *American Economic Review*, 86 (3), 562-584

Knight, F. (1921), *Risk, Uncertainty and Profit*, Houghton Mifflin: Boston.

Krugman, P. (1990), *Rethinking International Trade*, MIT Press: Cambridge (MA).

Laffont, J-J., P. Rey and J. Tirole (1998), "Network Competition I: Overview and Nondiscriminatory Pricing," *Rand Journal of Economics*, 29 (1), 1-37.

Lucas, R.E. (1978). "On the Size Distribution of Business Firms," *Bell Journal of Economics*, 9, 508-523.

Marris, R. and D. Mueller (1980), "The Corporation, Competition and the Invisible Hand," *Journal of Economic Literature*, 18 (1), 32-63.

MasColell, A., M. Whinston and J. Green (1995), *Microeconomic Theory*, Oxford University Press: Oxford.

McAfee, P. and J. McMillan (1995), "Informational Diseconomies of Scale," *Journal of Economics and Management Strategy*, 4 (3), 399-426.

Meyer, M., P. Milgrom and J. Roberts (1992), "Organizational Prospects, Influence Costs and Ownership Changes," *Journal of Economics and Management Strategy*, 1, 9-35.

Milgrom, P. and J. Roberts (1992), *Economics, Organization and Management*, Prentice-Hall: Englewood Cliffs.

Murphy, K., A. Shleifer and R. Vishny (1989), "Industrialization and the Big Push," *Journal of Political Economy*, 97 (5), 1003-1026.

Penrose, E. (1959), *The Theory of the Growth of the Firm*, Sharpe: New York.

Porter, M.E. (1980), *Competitive Strategy*, Free Press: New York.

Qian, Y. (1994), "Incentives and Loss of Control in an Optimal Hierarchy," *Review of Economic Studies*, 61 (3), 527-544.

Romer, P.R. (1994), "New Goods, Old Theory and the Welfare Costs of Trade Restrictions," *Journal of Development Economics*, 43, 5-38.

Salant, S., Switzer, S. and Reynolds, R. (1983), 'Losses from horizontal merger: The effects of an exogenous change in industry structure on Cournot-Nash equilibrium', *Quarterly Journal of Economics* 98(2), 185–99.

Salop, S. and J.E. Stiglitz (1977), "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersions," *Review of Economic Studies*, 44, 493-510.

Simon, H.A. and C.P. Bonini (1958), "The Size Distribution of Business Firms," *American Economic Review*, 48 (4), 607-617.

Stern, S. (1995), "Product Demand and Pharmaceutical Markets," *mimeo*., MIT.

Stigler, G.J. (1958), "The Economies of Scale," reprinted in *The Essence of Stigler*, K.R. Leube and T.G. Moore (eds.), Hoover Institution, Stanford, 1986, 25-45.

Sutton, J. (1991), *Sunk Costs and Market Structure*, MIT Press: Cambridge (MA).

Sutton, J. (1996), "Technology and Market Structure," *European Economic Review*, 40 (3-5), pp.511-530.

Sutton, J. (1997), "Gibrat's Legacy," *Journal of Economic Literature*, 35 (1), 40-59.

Sutton, J. (1998), *Technology and Market Structure*, MIT Press; Cambridge (MA).

Tirole, J. (1988), *The Theory of Industrial Organization*, MIT Press: Cambridge (MA).

Viner, J. (1931), "Cost Curves and Supply Curves," in *Zeitschrift f. Nationalokon* reprinted in G.J. Stigler and K.E Boulding (eds.), *AEA Readings in Price Theory*, Irwin: Homewood (IL), 1952, 198-232.

Von Ungern-Sternberg, T. (1988), "Monopolistic Competition and General Purpose Products," *Review of Economic Studies*, 55, 231-246.

Williamson, O.E. (1981), "The Modern Corporation: Origins, Evolution, Attributes," *Journal of Economic Literature*, 19 (4), 1537-1568.

Winter, S.G. (1984), "Schumpeterian Competition in Alternative Technological Regimes," *Journal of Economic Behavior and Organization*, 5, 287-320.