Why are Firms Hierarchical?

MARK CASSON

ABSTRACT *Firms are hierarchical because hierarchy facilitates both monitoring and control. Monitoring is readily explained in terms of agency costs, but control is not. Control relations between superiors and subordinates emerge naturally only when the superior has* decisive *information and the subordinate does not. When decisiveness does not naturally occur, it may nevertheless be imposed in order to reduce communication costs. The degree of decisiveness reflects the pattern of volatility in the firm's environment. The theory connects the volatility of the environment to the degree of hierarchy in the organisation. Recent changes in volatility can be used to explain contemporary demands for flatter organizational structures.*

Key words: Hierarchy; Decisiveness; Monitoring; Control.

JEL classification: L22.

1. Introduction

Hierarchy is an unpopular concept nowadays. Nobody likes having to take orders from someone else, and they also resent being checked up on to see that the orders have been obeyed. Orders are particularly objectionable when those who make them do not consult in advance with those who have to carry them out. Such orders often appear arbitrary, if not mistaken, with the result that morale in the lower ranks of organizations is frequently low.

But if hierarchies are so demoralizing, why are they so prevalent? Why do hierarchies not collapse under competitive pressure from more participatory organizational forms? There is currently considerable interest in introducing flatter organizational structures into firms (Kanter, 1989). But why did this movement not begin earlier? Is there, perhaps, an optimal degree of hierarchy which depends upon environmental conditions? If so, what factors determine this optimum, and what are the trade-offs involved? What environmental changes explain the current switch to less hierarchical structures? These are the main questions addressed in this paper.

To answer these questions is no simple matter, because the nature and rationale of hierarchy is still not fully understood. The literature on organizational behaviour tends to take the existence of hierarchy as self-evident (for a survey see Nystrom and Starbuck, 1981). It is mainly economists who have

Mark Casson, Professor of Economics, Department of Economics, University of Reading, Faculty of Letters and Social Sciences, PO Box 218, Whiteknights, Reading RG6 2AA, UK.

compared hierarchies with alternative organizational forms to see whether they are efficient or not. Transaction costs have been widely used for this purpose. Coase's (1937) analogy between a hierarchy and an internal market is well known, and so is the typology of alternative organizational forms that Williamson (1985) has developed from it.

While the transaction cost approach emphasizes the benefits of hierarchy, agency theory focuses on minimizing its bureaucratic costs (see Milgrom and Roberts, 1992; Wiggins, 1991). Hierarchy allows managerial employees to be monitored, and ensures that the monitors themselves can be monitored too. Monitoring provides the employer with the information needed to give incentive for managerial effort through a system of performance-related pay.

There is more to hierarchy than monitoring, though. The information flowing through a hierarchy is not just about how its members perform. It is also about the firm's environment, and about the strategies that are available for responding to it. This emphasis on the more general issue of information processing is reflected in the work of Simon (1957) and Marschak and Radner (1972), who analyse hierarchy as a decision-making mechanism which economizes on information costs. It is this more general approach, recently summarized by Radner (1992), which is developed here.

There is other work on hierarchies which is tangential to this paper: for example, Sah and Stiglitz (1986) on sequential screening, Aoki (1990) on horizontal *versus* vertical information flow and Marglin (1974) on the internal distribution of power. Unfortunately, though, there is insufficient space to explore the relevance of this literature here.

The answers to the questions raised above are developed in three steps. The first clarifies the concept of hierarchy as a pyramid of authority (Figure 1). Several issues emerge from this exercise: most are straightforward, but one is not. The difficult issue relates to the separation of policy-making and policy-implementation that is implicit in the notion of control. In particular, how can it be rational for the policy-maker to ignore apparently relevant information that the implementer holds, as so often happens in practice?

This issue can be resolved using the concept of decisiveness. The second step examines this concept and demonstrates its vital role. Decisiveness determines who controls whom within the hierarchy, and hence governs the distribution of

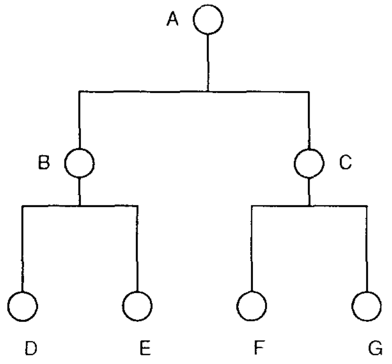

Figure 1. A typical schematic representation of hierarchy.

decision-making power. A high level of decisiveness supports a tall hierarchy, and a low level of decisiveness a flat one.

The level of decisiveness in turn reflects the environment of the firm—in particular, the pattern of external shocks. Different patterns of shocks generate different degrees of decisiveness and hence encourage different kinds of hierarchical structure. The final step applies this insight to explain recent trends in the organizational structure of firms.

2. The Concept of Hierarchy

Examination of Figure 1 shows that there are two distinct aspects of hierarchy. The first is the relation between superior and subordinate, which is the basic elemênt from which the hierarchy is assembled. The second is the pyramid structure which configures the relations of subordination between the different people involved.

The relationship between superior and subordinate is fundamentally asymmetrical, in the sense that their roles cannot be reversed. There are, however, two distinct aspects to this relation which in popular discussion are often confused: namely monitoring and control. As noted above, the monitoring function tends to be emphasized in the economics literature, almost to the exclusion of control (Radner, 1992). In practice both are present: what is monitored tends to be the effectiveness of managers in exercising control (Miller, 1992).

The pyramid structure has three main properties (see Table 1):

- unity of command (one person at the top);
- span of responsibility (each superior has several subordinates); and
- stacking (in the middle ranks, one person's subordinate is another person's superior).

These properties are fairly easy to rationalize where monitoring is concerned—see the left hand column of the table. Unity of command is explained by

Table 1. Key properties of a hierarchy

	Nature of relation	
Structural property	Monitoring	Control
Unity of command	Owner(s) wishes to hold a single person ultimately accountable for the performance of the firm as a whole.	The problem of determining the firm's optimal plan must ultimately be endorsed by a single 'brain'.
Spanning	Each monitor has the capacity to observe several people—an economy of scale in observation.	Applying a division of labour to problem-solving generates a series of sub-problems. The problem as a whole is solved by reconciling and then synthesizing the solutions to the sub-problems.
Stacking	Monitors can themselves be monitored.	Sub-problems can in turn be further decomposed.

the owner's need to hold a single person ultimately responsible for the performance of the firm. This simplifies matters by reducing the ownership function to hiring the appropriate chief executive and replacing him when required (Knight, 1921).

The span of responsibility is explained by economies of scale in monitoring. The indivisibility of a specialized monitor creates a capacity to observe several people at once—the exact number depending on the architecture of the workplace and the monitoring technology employed. A further economy arises from the fact that a wide span of control generates a wealth of comparative information on subordinates which can be used to establish a reliable yardstick for individual performance.

The principle of stacking is explained by the fact that monitors can themselves be monitored too. The application of this principle is restricted, however, by the fact that monitoring itself is not so easy to monitor as the simple manual work that is typically carried out at the lowest level of the hierarchy.

The pyramid structure is more difficult to rationalize where control is concerned—the remarks in the right hand column of the table are more tentative because of this. Intuitively, a pyramid of control emerges when a complex problem, for which one person is ultimately responsible, is broken down into a series of subproblems whose solution is delegated to different people. Unfortunately, though, the solution of each subproblem will normally be conditional on the solutions of other subproblems. Unless there is extensive lateral consultation, therefore, it may be difficult, if not impossible, to match up the solutions when they come to be synthesized.

Lateral consultation is, however, difficult to reconcile with a pyramid in which those at the top make policy and those at the bottom are simply instructed to carry it out. Decisiveness holds the key as to whether or not consultation will occur. Differences in decisiveness mean that some problems have a logical structure which supports solution without consultation and some do not. Furthermore, it is often worthwhile to inhibit consultation, even where logic would require it, simply because the expected gain from an improved decision does not outweigh the additional communication cost involved. Thus some problems are naturally hierarchical, some favour hierarchy when communication costs are high, and some will not support hierarchy at all.

The contingent nature of this conclusion probably explains why hierarchies are often explained in terms of monitoring even though control is their most conspicuous feature. Monitoring predicts a fairly standardized pyramid—symmetrically shaped with an even pattern of steps—whereas control predicts a variety of forms—some tall, some flat, some symmetric and some asymmetric, and some having very different numbers of steps at each stage.

3. The Concept of Decisiveness

The concept of decisiveness confronts head-on the issue of why information possessed by a subordinate might be ignored by a superior. The natural context in which to develop the concept is a firm whose owner hires two employees to formulate and implement a production plan. To begin with, the situation is fully symmetrical, in the sense that both employees possess potentially relevant information of their own, and share responsibilities for implementation. Yet out of this may emerge a situation in which one of them is empowered by the owner to give orders to the other. Moreover, the superior need not consult the subordinate beforehand.

Particular attention is given to the question of when this control relation will emerge and when it will not. This in turn requires that considerable care is taken in modelling the environmental shocks that impinge upon the firm. To focus on this issue it is useful to eliminate further complications. Thus while the control relation is considered in detail the pyramid structure built out of this relation is not. Because only two people are involved, unity of command is trivially satisfied, but the span of control is restricted to one. Moreover, no stacking of control relations in involved.

Each period the owner seeks a plan which will maximize the expected profit of the firm. This profit is measured net of information costs. The relevant information pertains to two environmental factors, each of which affects the relative profitability of alternative production plans. Each factor can be observed by one of the employees but not the other. This is because information on each factor is obtained as a byproduct of implementing the previous period's production plan. The division of labour in implementation generates a division of labour in information gathering. Thus information is distributed between the employees and can be concentrated only through communication.

In principle the two employees can communicate either with each other or with a third party such as the owner of the firm. However, to limit the number of decision-makers to two, only the first possibility is considered here—it is assumed that the owner of the firm is too busy with other things to actively participate in the management of the firm. This means that the control relation develops between the employees themselves, and not between the employees and the owner (although in practice, of course, these relations often occur together).

Information on one particular factor may prove decisive for one aspect of the plan. It is decisive in the sense that if the value of the factor is high then one strategy should be pursued, while if the value is low another strategy should be pursued, independently of what the value of the other factors turns out to be. Thus if employee A has direct access to decisive information, then he can act without consulting employee B. But B, who lacks decisive information, may have to consult A first. Thus people with decisive information can act without consulting others even though others are obliged to consult with them.

Asymmetry in consultation can lead to asymmetry in order-giving. There are two main reasons for this. The first is that it may be more relevant for B, when consulting A, to know what A has decided, than to know what A knew when making his decision. Thus in the keynote example below it is more relevant for a production manager B to know what sales target the marketing manager A has fixed in the light of market research than to know what this market research actually revealed.

The second reason is that the decision may be easier to communicate than the information on which it is based. Thus even if B needed to know what A's research had discovered, the results of this research may be more difficult to communicate than the decision that A has reached. This is because information is often more tacit than the decisions which are based upon it (Polanyi, 1964; Foss, 1993). Thus after allowing for the differential cost of communicating the results of the market research and the marketing plan, it may be better to communicate just the marketing plan even though B would ideally like to know the results of the market research instead.

The thrust of the argument, therefore, is that hierarchy emerges when people with access to decisive information, which others need to consult, inform these other people of their decisions rather than of the information itself. Because their information is decisive they do not need to consult other people. Because their information is tacit it is easier for them to communicate their decision rather than their information. And provided their decision is an adequate proxy for the information they possess, it is sufficient for other decision-makers to make do with the decision information they are offered.

The practical application of this argument requires, though, that real-world problems have a logical structure that supports hierarchical division of decision-making along these lines. Substantiating this point is the major rationale for the formal model presented below. A simple way of interpreting this model is to say that it highlights a distinction between strategy formulation and tactical implementation which naturally occurs in some decision problems. Decisive information supports a strategic decision which can be made without reference to the particular tactics which will be used to make it work. The non-decisive information supports a tactical decision as to how the chosen strategy will be implemented. The implementer must know what the strategy is in order to choose the correct tactics, but the strategist does not need to know what the tactic will be before selecting the strategy.

4. The Formal Model

Consider an entrepreneur who has set up a firm to create a market for a new consumer good. He hires two assistants—a marketing assistant to distribute the good to consumers and a production assistant to recruit and pay the workers. The owner recognizes two sources of volatility in the environment: consumer fashion, and technological variability. These factors generate discrete changes in demand and supply.

Each period the state of demand is either buoyant ($s_1 = 1$) or depressed ($s_1 = 0$). The firm can respond by setting either a high level of output ($x_1 = 1$) or a low level of output ($x_1 = 0$). Two production techniques are available. The first ($x_2 = 0$) is a labour-intensive technique which involves low fixed costs and high variable costs, while the other ($x_2 = 1$) is a capital-intensive technique which involves lower variable costs. In a favourable environment ($s_2 = 1$) the fixed cost of the capital-intensive technique is the same as that of the labour-intensive technique, but otherwise ($s_2 = 0$) it is much higher. The fixed costs are recurrent costs that are incurred each period.

In practice, of course, demand tends to be much more volatile than production technology in the short run, although technology can be quite volatile in the long run. This formulation of the model has been chosen for its analogy with the conventional theory of the firm, whose static nature tends to obscure such differences. In practical applications of the model it might be more helpful to identify production shocks with fluctuations in raw material supply.

Permuting the binary states of the two factors s_1 and s_2 generates four possible outcomes. The value of each of these outcomes depends on the values of the instruments x_1 and x_2. The entrepreneur knows the profit generated by each of the 16 possible strategy–state combinations—in other words, he knows the profit–function $\pi(x_1, x_2, s_1, s_2)$. When s_1 and s_2 are both known at the start of each period then the entrepreneur can specify an optimal decision rule relating x_1 to s_1 and s_2 and another rule relating x_2 to s_1 and s_2.

It is assumed that the implementation of each rule is delegated to the appropriate assistant. The marketing assistant fixes the output x_1 and the production assistant chooses the technique x_2. The key influence on hierarchical structure is the cost of investigating the states s_1 and s_2, and communicating this information to the relevant people. Only the marketing assistant can readily investigate the state of demand, because this information is generated as a byproduct of his work in distributing the product. For similar reasons only the production assistant can readily investigate the state of technology. To begin with, investigation costs are ignored.

In general the output decision x_1, made by the marketing assistant, will depend not only on the state of demand s_1, which he has discovered for himself, but on the state of technology s_2, which he must learn from the production assistant. The state of technology is tacit information, however, which incurs a cost of communication $m > 0$. If the production assistant were simply to tell the marketing assistant what technique he has chosen then this would be explicit information and the cost of communication would be zero. But the production assistant cannot do this when his own decision x_2 depends not only on the state of technology he has discovered for himself but on the state of demand s_1 which he must learn from the marketing assistant instead. Thus tacit information on the state of demand must be communicated to the production assistant at an additional cost. The total cost of communication between the assistants is thus $2m$.

The communication of tacit information can be avoided, and the cost of communication consequently eliminated, if one of the items of information s_1, s_2 is decisive. The state s_1 is said to be decisive for x_1 if x_1 varies according to s_1 only, and is independent of s_2. It should be noted that this is a property of an optimal decision rule, extracted from a pair of rules derived by the entrepeneur. This means that the value of x_2 is being optimized at the same time, so that each value of x_1 chosen is associated with a best response from x_2.

The decisiveness of s_1 in this case signifies that the state of demand is so crucial that it alone determines the output decision. The typical case is where depressed demand calls for low output and buoyant demand for high output, independently of the fixed costs incurred by the capital-intensive technique. Conversely, the state s_2 is decisive for x_2 if x_2 varies according to s_2 only, and is independent of s_1. The typical case is where high capital cost stimulate a switch to labour-intensive production independently of the state of demand. Note that decisiveness excludes the case where the strategy is constant.

Other patterns of decisiveness are possible. Thus s_1 could be decisive for x_2: production would be capital-intensive, for instance, if and only if demand was buoyant. Alternatively, s_2 could be decisive for x_1: output would be high, for example, only if capital costs were low. These cases are highly improbable. More plausible are cases where certain patterns of decisiveness coexist together: thus if revenues were very sensitive to demand, but capital costs were never a high proportion of total cost, s_1 might be decisive for both x_1 and x_2.

It is easy to see how decisiveness eliminates communication costs. Suppose, for example, that s_1 is decisive for x_1. Then the marketing assistant can determine output without reference to the level of capital costs. Because each value of x_1 is generated by a unique value of s_1, the entrepreneur can express the production assistant's decision rule in terms of x_1 rather than s_1. Thus if the marketing assistant tells the production assistant what output he has decided the production assistant can combine this information with his own information about the state of technology to determine the appropriate technique x_2. In effect, the marketing assistant gives an order specifying the level of output and the production assistant uses his own information to implement the order by an appropriate choice of technique. A single order has replaced a two-way flow of information; since the order is explicit, while the information it replaces is tacit, communication costs have effectively been eliminated.

Decisiveness can be a conditional concept too. For example, $s_1 = 0$ might imply $x_1 = 0$, whereas $s_1 = 1$ would not be sufficient to decide whether $x_1 = 0$ or $x_1 = 1$. This is the case where depressed demand definitely calls for low output, but buoyant demand only calls for high output when capital costs are low. Conditional decisiveness still affords communication economies, but the expected value of the savings reflects, of course, the frequency with which the condition is satisfied.

5. The Marketing-led Firm

The preceding discussion suggests that four main patterns of decisiveness are likely to be observed in practice. In the first, the state of demand is decisive for output, but the state of technology is not decisive for the choice of technique. This creates a 'marketing led' firm in which the marketing assistant tells the production assistant what output should be produced. In the second case, the state of technology is decisive for choice of technique but the state of demand is not decisive for output. This creates a 'production led' firm in which the production assistant tells the marketing assistant how the output will be produced and leaves the marketing assistant to decide how much he requires. In the third case both kinds of decisiveness prevail and the firm becomes functionally separated: the marketing assistant determines the output without reference to costs and the production assistant determines technique without reference to demand. Finally, there may be no decisiveness at all. The marketing assistant consults the production assistant about technology and the production assistant consults the marketing assistant about demand. This generates the 'consultative' firm.

Casual empiricism suggests that the marketing led firm is a particularly common type. It is fairly easy to construct a simple example which suggests that this may indeed be so. For this purpose it is sufficient to specify a profit function and examine how decisiveness varies according to its parameter values.

Suppose, therefore, that the firm is engaged in batch production, and has a capacity of two batches each period. If it produces just one batch ($x_1 = 0$) then in a depressed market ($s_1 = 0$) the batch will sell for $P_0 > 0$. If it produces two batches ($x_1 = 1$) then each batch will sell for P_1 ($0 < P_1 \leq P_0$), generating a revenue $2P_1$. In a buoyant market ($s_1 = 1$) each batch produced will command a price premium $r > 0$; thus revenue will $P_0 + r$ if one batch is produced and $2P_1 + 2r$ if two batches are produced. Labour-intensive production incurs a constant cost per batch c_1 ($0 < c_1 < c_0$). If technology is favourable ($s_2 = 1$) then a fixed cost $f_1 \geq 0$ is incurred, but if it is unfavourable ($s_2 = 0$) then a fixed cost $f_1 + f_2$ ($f_2 > 0$) must be paid instead. This generates the profit function shown in Table 2.

To make the choice of strategy non-trivial it is useful to restrict the parameters so that both factors are relevant. In other words, in at least one state of

Table 2. Example of a profit function

	s_2:	0		1	
s_1	x_1 x_2:	0	1	0	1
0	0	$P_0 - c_0$	$P_0 - c_1 - f_1 - f_2$	$P_0 - c_0$	$P_0 - c_1 - f_1$
	1	$2P_1 - 2c_0$	$2P_1 - 2c_1 - f_1 - f_2$	$2P_1 - 2c_0$	$2P_1 - 2c_1 - f_1$
1	0	$P_0 + r - c_0$	$P_0 + r - c_1 - f_1 - f_2$	$P_0 + r - c_0$	$P_0 + r - c_1 - f_1$
	1	$2P_1 + 2r - 2c_0$	$2P_1 + 2r - 2c_1 - f_1 - f_2$	$2P_1 + 2r - 2c_0$	$2P_1 + 2r - 2c_1 - f_1$

Note: To produce economically meaningful results, $s_1 = 0$, $s_2 = 0$ should imply $x_1 = 0$, $x_2 = 0$, while $s_1 = 1$, $s_2 = 1$ should imply $x_1 = 1$, $x_2 = 1$. These conditions are satisfied when $f_1 + f_2 > c_0 - c_1 > (f_1/2)$ and $2P_1 + r - c_1 > P_0 > 2P_1 - c_0$. If these conditions are not satisfied then only one of the factors is relevant to the firm's decisions.

technology an increase in demand will induce an increase in output, and in at least one state of demand an improvement in technology will induce a switch to the capital-intensive technique. Furthermore, because costs are independent of price, the influence of demand on the choice of technique is always mediated by output. This means that a change in demand can change the technique only if it changes the level of output too. These restrictions drastically reduce the number of different cases that need to be considered, and allow most of the important results to be extracted from the model using just a couple of examples.

Suppose, for instance, that the price for which a batch can be sold in a depressed market falls from $P_0 = 4$ to $P_1 = 3$ when an additional batch is produced, and that the premium obtained in a buoyant market is $r = 1$. Let the marginal cost of labour-intensive production be $c_0 = 3.25$ and the marginal cost of capital-intensive production be $c_1 = 2.25$. If technology is favourable then capital-intensive production incurs no fixed costs ($f_1 = 0$), but if it is unfavourable the fixed costs are $f_2 = 1.5$. This generates the profit values shown in Table 3.

The table is divided into four blocks, each block corresponding to the four strategic responses that can be made to one of the four possible combinations of states. Within each block the optimal strategic response is determined by the row and column in which the highest profit appears. The optimal responses are identified by an asterisk. All of these responses lie in either the top row or the bottom row of the table. This indicates that, with this pattern of profit, the state of demand s_1 is decisive for output x_1. However, the column in which the optimal response appears varies according to the row, which means that s_2 is not decisive for x_2. This case therefore corresponds to the marketing-led firm.

Changing one or more of the parameters is liable to affect the decisiveness. For example, if the fixed cost of capital-intensive production under favourable conditions rises to $f_1 = 1.5$, while f_2 remains the same, then the profits change to those shown in Table 4. The second and fourth columns are each reduced by 1.5, with the result that labour-intensive production is preferred when either demand is low or fixed costs are high, and low output is preferred to high output when demand is high and fixed costs are high as well. As a result, demand is no longer decisive for output, and consultation is appropriate.

This is, in fact, the paradigm case in which an intermediating firm, responsive to both demand and technology, is faced with non-decisive information. It is not

Table 3. A numerical example

	s_2:	0		1	
s_1	x_1 x_2:	0	1	0	1
0	0	0.75*	0.25	0.75	1.75*
	1	−0.5	0.0	−0.5	1.5
1	0	1.75	1.25	1.75	2.75
	1	1.5	2.0*	1.5	3.5*

Note: An asterisk indicates the profit-maximizing response to each of the four possible combinations of states. These responses generate the following decision rules:

$$x_1 = x_1(s_1) = \begin{cases} 0 & \text{if } s_1 = 0 \\ 1 & \text{if } s_1 = 1 \end{cases}$$

$$x_2 = x_2(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 = s_2 = 0 \\ 1 & \text{if either } s_1 = 1 \text{ or } s_2 = 1 \end{cases}$$

Since the first rule implies $x_1 = s_1$, the second rule can be expressed in the form

$$x_2 = x_2(x_1, s_2)$$

This shows that the decisiveness of s_1 for x_1 can be exploited to eliminate tacit communication altogether, since x_2 can be set using information on x_1 rather than s_1.

Table 4. A numerical example with higher capital costs

	s_2:	0		1	
s_1	x_1 x_2:	0	1	0	1
0	0	0.75*	−1.25	0.75*	0.25
	1	−0.5	−1.5	−0.5	0.0
1	0	1.75*	−0.25	1.75	1.25
	1	1.5	0.5	1.5	2.0*

Note: An asterisk indicates the profit-maximizing response to each of the four possible combinations of states. These responses generate the following decision rules:

$$x_1 = x_1(s_1, s_2) = \begin{cases} 0 & \text{if either } s_1 = 0 \text{ or } s_1 = 1 \text{ and } s_2 = 0 \\ 1 & \text{if } s_1 = 1 \text{ and } s_2 = 1 \end{cases}$$

$$x_2 = x_2(s_1, s_2) = \begin{cases} 0 & \text{if either } s_2 = 0 \text{ or } s_2 = 1 \text{ and } s_1 = 0 \\ 1 & \text{if } s_1 = 1 \text{ and } s_2 = 1 \end{cases}$$

This generates a consultative approach to decision-making.

the only case that can arise, but economic intuition suggests that it is the most likely case. Furthermore the other economically-meaningful cases of non-decisiveness are logically isomorphic to it, in the sense that they interchange the role of 'good' and 'bad' conditions, or of demand and technology, but leave the basic structure of the argument unchanged.

The most instructive way to analyse this case is to focus on the two parameters f_2 and r which mediate the impact of the volatile factors on the costs and revenues of the firm. The comparative statics of the parameters f_2 and r are fairly straightforward, and because there are just two of them the results can be presented easily in graphical terms.

Holding the five other parameters P_0, P_1, c_0, c_1, f_1 at the values assumed in Table 4 generates the profit structure shown in Table 5. The anchor point for analysing Table 5 is the top right-hand quadrant, where the labour-intensive low-output strategy $x_1 = 0$, $x_2 = 0$ is identified as the optimal response to depressed demand and favourable technology, $s_1 = 0$, $s_2 = 1$, independently of the values of f_2 and r. This means that if s_1 is to be decisive for x_1 then $x_1 = 0$ must be the optimal response to $s_1 = 0$, $s_2 = 0$. The top left-hand block of figures shows that this is guaranteed for any positive value of f_2.

It is also necessary for decisiveness that $x_1 = 1$ be the optimal response to $s_1 = 1$. Looking at the bottom right-hand block, it can be seen that for a positive premium r the optimal response to $s_1 = 1$, $s_2 = 1$ is either $x_1 = 0$, $x_2 = 0$ or $x_1 = 1$, $x_2 = 1$. The latter will be preferred, as decisiveness requires, whenever $r \geq 0.75$. The most serious constraint on decisiveness, however, occurs in the bottom left-hand block, where it is necessary to ensure that high output ($x_1 = 1$) will be set even if technology is unfavourable ($s_2 = 0$). It is readily established that for low output, capital-intensive production ($x_1 = 0$, $x_2 = 1$) is less efficient than labour-intensive production ($x_1 = 0$, $x_2 = 0$) for any positive combination of f_2 and r, but the relative merits of the former compared to the high-output strategies need to be examined using a series of pairwise comparisons between them.

The results of this exercise are shown in Figure 2. Demand is decisive for output if either $r > 1.25$ or $r > 0.75 + f_2$—in other words, if the price premium is sufficiently large, either in absolute terms, or relative to the fixed cost premium. If, in addition, the fixed cost premium is less than 0.5 then demand is decisive for the choice of technique as well: buoyant demand requires output produced using the capital-intensive technique and depressed demand calls for a small output produced uisng the labour-intensive technique.

It should be noted that even when absolute decisiveness does not occur, conditional decisiveness is present: low demand is always decisive for low output and for the labour-intensive technique. Indeed, when the price premium drops below $r = 0.75$ the firm's strategy becomes completely trivial—it always pays to produce high output using the labour-intensive technique.

Finally, it should be noted that demand-decisiveness is the only kind of decisiveness that appears in this case. Decisiveness based on technology never occurs. This leads support to the idea that decisiveness, when it occurs, tends to

Table 5. Analysis of the effect of the price premium r and the fixed cost premium f_2 on the pattern of decisiveness

		s_2:	0		1	
s_1	x_1	x_2:	0	1	0	1
0	0		0.75	$0.25 - f_2$	0.75*	0.25
	1		-0.5	$-f_2$	-0.5	0.0
1	0		$0.75 + r$	$-0.25 - f_2 + r$	$0.75 + r$	$0.25 + r$
	1		$-0.5 + 2r$	$-f_2 + 2r$	$-0.5 + 2r$	$2r$

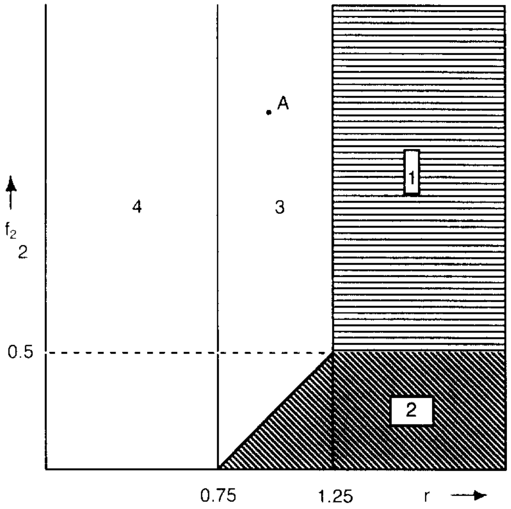

Figure 2. Influence of price premium and fixed cost on decisiveness. *Note:* 1. Marketing-led-firm; demand is decisive for output. 2. Marketing-dominated firm: demand is decisive for both output and production technique. 3. Consultative firm: no decisiveness. 4. Low-output labour-intensive firm. A. Point corresponding to data in Table 3.

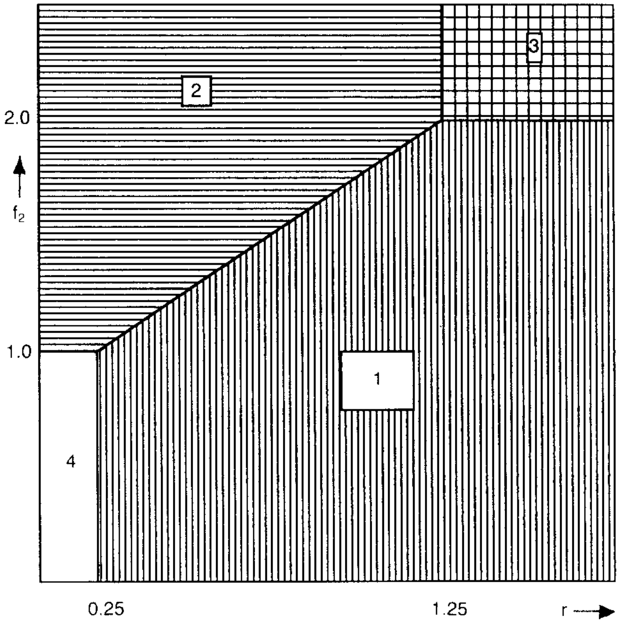

Figure 3. Influence of price premium and fixed cost on decisiveness when capital costs are low. *Note:* 1. Marketing-led-firm; only demand is decisive. 2. Production-led-firm: only technology is decisive. 3. Separable firm: both are decisive. 4. High-output capital-intensive firm.

generate the marketing-led firm. However, it is fairly obvious why, in this example, technology is not decisive; it is because the fixed cost of capital-intensive production f_1 is relatively high, so that labour-intensive production is normally the preferred technique. If f_1 is reduced to zero, as in Table 3, then a rather different picture emerges. Figure 3 shows the various possibilities associated with different premia when $f_1 = 0$. Demand is decisive for output if $r > 0.25$ and $f < 0.75 + r$ $(0.25 \leq r \leq 1.25)$. Technology is decisive for technique if $f_2 > 1$ and $f_2 > 0.75 + r$ $(0.25 \leq r \leq 1.25)$. Both conditions are satisfied—so that the decisions are functionally separated—if $r > 1.25$ and $f_2 > 2$, whilst neither condition is satisfied if $r < 0.25$ and $f_2 < 1$.

It can be seen that even here the conditions for demand to be decisive are less strict than the conditions for technology to be decisive. The former requires a price premium of only 0.25, which generates, at most, additional revenue of 0.5, while the latter requires a fixed cost premium of at least unity. Thus a much larger shock to profits is required to make technology decisive than to make demand decisive instead. This suggests that the marketing-led firm is indeed the normal consequence of decisive decision-making. Of course, this result is conditional on the parameter values which have been assumed. Its practical significance hinges on the view that these values (or at least their relative magnitudes) are representative of those in real world situations.

6. How Hierarchy Suppresses Communication

The preceding discussion has focused on the way in which decisiveness can arise naturally. It is, however, quite possible for decisiveness to be imposed, in the sense that decision rules can be modified to reduce the need for tacit communication. Even though tacit communication may be required to guarantee a correct decision, it may be advantageous to dispense with it in order to economize on communication costs. This means substituting a rule which is technically incorrect, in that it fails to condition strategic choice on a relevant environmental factor, in order to avoid the cost of communicating information about this factor to the decision-maker. The net benefit of this strategy depends on the magnitude of the communication cost that is avoided, and the probability that a strategic error will occur.

To assess the probability of strategic error the owner needs to form subjective probabilities about the states of the environmental factors. Let p_1 be the probability that $s_1 = 1$ and p_2 the probability that $s_2 = 1$ $(0 \leq p_1, p_2 \leq 1)$. It is assumed that these probabilities are independent, so that the volatile factors are 'orthogonal', i.e. uncorrelated. These are subjective probabilities which could, in principle, be updated over time as the owner accumulates experience. Revision of subjective probabilities may lead to changes in hierarchical structure but to keep the analysis simple, the effects of learning will not be considered here.

It is assumed that the owner is risk-neutral, and so maximises expected profit, v, where profit is now measured net of information cost, i:

$$v = \sum_{s_1=0}^{1} \sum_{s_2=0}^{1} p(s_1, s_2)\pi(x_1, x_2, s_1, s_2) - i$$

where

$$p(0, 0) = (1 - p_1)(1 - p_2) \qquad p(0, 1) = (1 - p_1)p_2$$

$$p(1, 0) = p_1(1 - p_2) \qquad p(1, 1) = p_1 p_2$$

Expected profit is maximized by selecting the appropriate pair of decision rules. Each decision rule in a pair relates one of the strategic variables x_1, x_2 to the two factor values s_1, s_2. A set of rules can assign any one of four strategic responses to each of the four combinations of states, so in principle there are $4^4 = 256$ different rules to choose from. Fortunately the features which give economic meaning to a problem usually make it possible to eliminate all but a few of the possible rules at an early stage of the solution process.

When communication costs are minimized, some rules will require tacit communication and others will not. Suppose for the moment that costs are incurred in keeping channels of tacit communication ready for use and are independent of the frequency with which they are utilized. Then the communication cost associated with each rule is independent of the probabilities p_1, p_2. Indexing rules by k gives the constraints

$$x_1 = x_1(s_1, s_2, k); \qquad x_2 = x_2(s_1, s_2, k); \qquad i = i(k)$$

and v is then maximized by selecting k.

Consider again the numerical example presented in Table 4. It was shown that in the absence of communication costs the optimal rule is to produce low output using the labour-intensive technique ($x_1 = 0$, $x_2 = 0$), unless buoyant demand coincides with favourable technology ($s_1 = 1$, $s_2 = 1$), in which case high output is produced using the capital-intensive technique. There is no point in considering alteratives to this rule unless, when communication costs are positive, they involve less tacit communication, which means, in turn, that they are more decisive than the original optimum.

'Perverse' decision rules, which make demand decisive for choice of technique and technology decisive for output, are readily seen to be inferior to other rules, and is also easy to see how different rules for output and for technique should be paired up. As a result, search quickly converges on three alternatives to the original optimum.

Rule 1 makes demand decisive for output, $x_1 = s_1$, and sets technique in the same way as before. It simply suppresses the communication of technology to the marketing assistant, in other words.

Rule 2 makes technology decisive for technique, $x_2 = s_2$, and sets output in the same way as the original rule. It suppresses the communication of demand information to the production assistant instead.

Rule 3 eliminates all reference to the environment, and simply fixes output at a low level and commits the firm to the labour-intensive technique. This not only suppresses tacit communication but eliminates the need to investigate the environment as well. It also trivialises the decision rules. These additional economies are not considered at this stage, however, but only later. The reason why rule 3 can compete with the other more sophisticated rules is that it only makes an error in the case where buoyant demand and favourable technology coincide. Indeed it avoids errors made by the others. It correctly selects low output when demand is high but technology unfavourable, whereas rule 1 selects high output instead. It also correctly selects the labour-intensive technique when demand is depressed and technology unfavourable, whereas rule 2 selects the capital-intensive technique instead.

If the cost of one-way tacit communication is fairly high, $m = 0.05$, so that decisiveness saves 0.10 on two-way communication, then the original rule (rule 0) is no longer viable whatever the probabilities of various states of the environment

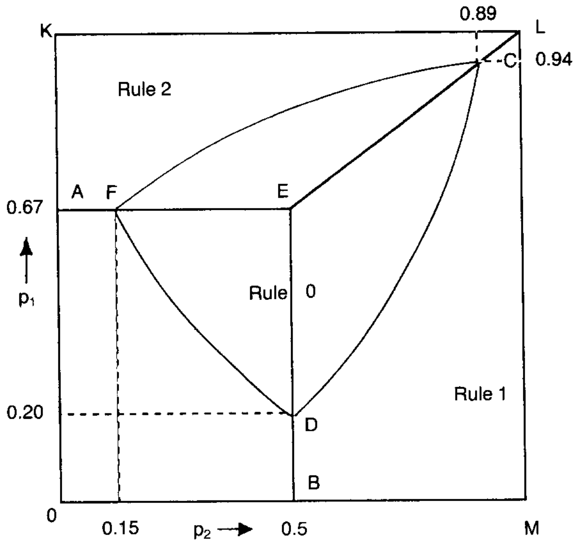

Figure 4. Use of decisiveness to economize on tacit communication. *Note:* Equations of segments. AE: $p_1 = 0.67$; BE: $p_2 = 0.5$; CE: $p_1 = 2p_2(1 + p_2)$; FC: $p_1 = 1 - (0.05/p_2)$; FD: $p_1 = 0.1p_2$; CD: $p_1 = 0.1/(1 - p_2)$.

happen to be. Figure 4 identifies three regimes in $p_1 - p_2$ space, each corresponding to a particular form of decisiveness which works better than the original rule. When demand is likely to be depressed and technology favourable (p_1 is low and p_2 is high) then it pays to make demand decisive for output, since the probability of an error, caused by setting output high when demand is high but technology is unfavourable, is relatively low. Thus in the region BELM rule 1 is preferred.

When demand is fairly certain to be buoyant (p_1 is high) it pays to make technology decisive for technique. The probability of an error caused by selecting the capital-intensive technique when demand is depressed and technology unfavourable is relatively low. Thus rule 2 is preferred in the region AKLE at the top of the figure.

Rule 3 is preferred when demand is likely to be depressed and technology unfavourable (the region OAEB) because then the conditions inducing error—namely buoyant demand and favourable technology—are unlikely to occur.

The three regions fill up the unit square entirely, which shows that whatever the subjective probabilities held by the owner, one of these rules is always better than the original one.

As tacit communication costs fall, of course, so the original rule becomes more attractive. Figure 1 also illustrates the solution that emerges when m falls to 0.0125. Rule 0 becomes attractive in all those indecisive situations where the subjective probabilities are close to neither zero or unity. This is equilvalent to the subjective variance of both volatile factors being relatively high. Rule 0 dominates the centre of the figure, occupying the curved triangular region FCD. As tacit communication costs rise and fall, so the triangular region expands away, from or contracts towards, the point of intersection E.

7. Why are Hierarchies Procedural?

Up to this point it has been implicitly assumed that two-way communication occurs simultaneously. In many cases this is a perfectly reasonable assumption: sequential communication would lead to unacceptable delays in making decisions. It cannot be ignored, however, that where a sequential procedure is feasible, significant economies may be available. The exploitation of these economies can have a major impact on organization behaviour, for life in the organization then becomes procedural.

Carrying out activities sequentially can afford better use of indivisible resources. For example, if there is a single channel of communication which can only operate in one direction at a time then communication between the marketing assistant and the production assistant must be a sequential affair. Furthermore tacit communication made in the first step may permit an order to be substituted for tacit communication in the second step. Thus the single channel can be set up to carry tacit information in only one direction rather than two.

Consider, for example, the implementation of rule 0—the rule that was derived from the data in Table 4. If communication is simultaneous then each assistant must supply tacit information to the other. But if communication is sequential then the production assistant can first communicate technology to the marketing assistant, and the marketing assistant can then give an order concerning the output. The production assistant simply follows a rule which says that low output must always be produced using the labour-intensive technique. Alternatively, the marketing assistant can communicate demand to the production assistant, who then gives an order concerning the technique. The marketing assistant then responds by setting low output if the labour-intensive technique is chosen, and high output if the capital-intensive technique is chosen instead.

Formally, what is happening is that the original pair of rules which related x_1 and x_2 to s_1 and s_2 is being replaced by one of two alternative pairs; either

$$x_1 = x_1(s_1, s_2), \qquad x_2 = x_1$$

or

$$x_2 = x_2(s_1, s_2), \qquad x_1 = x_2$$

respectively. Which of these pairs is chosen depends on the relative cost of tacit communication in each direction. If technology is easier to communicate than the state of demand then the production assistant will report to the marketing assistant, while if demand is easier to communicate than technology then it will be done the other way round.

The use of orders to exploit economies of sequential decision-making generates a rather different hierarchical pattern than does the exploitation of decisiveness. Decisiveness allows the person who gives the order to do so without consulting the recipient of the order, whereas sequential communication in non-decisive situations requires prior consultation with them. Decisiveness therefore captures the situation in which the superior takes decisions without ascertaining how the subordinate is likely to carry them out, whereas the sequential principle captures the situation where the superior requires the subordinate to report on this matter first.

8. Management by Exception

It is now appropriate to relax the assumption that the costs of communication are incurred only in keeping channels of tacit communication open, and to take account of costs incurred by each tacit message that is sent as well. When costs are related to the number of messages a new economy can be generated by designating one of the states of each factor as a normal state which does not need to be reported to the other party (Marschak and Radner, 1972:206). Each potential recipient of tacit information assumes that in the absence of communication the factor concerned is in its normal state. Typically it is the most probable state that is designated as the normal one—not only is this implicit in the concept of normality but, more saliently, it offers the greatest expected savings in message costs. States which are not normal are designated exceptional, and communication is thus concentrated on reporting exceptional situations.

Although the principle of suppressing normal information applies first and foremost to tacit information, it can be applied to the suppression of orders too. The suppression of both normal information and normal orders generates a distinctive pattern of hierarchical communication.

Consider, for example, the implementation of the rule 0 derived above. Costs are already being reduced by sequential communication, it is assumed. To begin with it is assumed that the production assistant reports to the marketing assistant rather than the other way round. The most probable state of technology, say the unfavourable state $s_2 = 0$, is identified as normal, and tacit information about the less probable, but more favourable, state $s_2 = 1$ is communicated only when it occurs. Likewise the most probable order—namely low output $x_1 = 0$—will also be identified as normal and consequently suppressed. The procedure is then for the production assistant to inform the marketing assistant only when an opportunity arises to exploit favourable technology—namely when $s_2 = 1$. If the production assistant receives no reply then he knows that the marketing assistant has decided not to raise output to take advantage of the situation—this is presumably because the marketing assistant has ascertained that demand is depressed. If he receives a reply, however, then he will be told that output is to be increased, and he knows to switch to the capital-intensive technique instead. This means that technological opportunity is going to be exploited after all.

Other examples can easily be constructed. The most obvious change is to make favourable technology the normal case. It then becomes problems rather than opportunities that are the subject of communication. So far as the marketing assistant is concerned 'no news is good news' in this case.

Formally, let m_1 be the cost of communicating exceptional demand information and m_2 the cost of communicating exceptional technology information. Let $q_1 = \min\{p_1, 1 - p_1\}$ be the probability that exceptional demand information (which, by construction, relates to the least probable state) needs to be communicated. Similarly $q_2 = \min\{p_2, 1 - p_2\}$ is the probability that exceptional technology information needs to be communicated. Then the expected cost of communication when the marketing assistant reports to the production assistant is $q_1 m_1$ compared to $q_2 m_2$ when the production assistant reports to the marketing assistant instead. Thus production becomes subordinate to marketing if $q_1 m_1 > q_2 m_2$.

This condition seems likely to be satisfied in practice. Demand factors tend to be more subjective and impressionistic, and therefore harder to communicate than production factors—although production factors may, of course, be difficult to explain to a marketing assistant who lacks a technological training. Demand factors are also likely to be more volatile: buoyant and depressed demand conditions are likely to alternate with greater frequency than good and bad technology conditions. This reinforces the case for communicating technology rather than demand.

Although the principle of suppressing normal information is very useful, it has a danger—namely that the channel of communication may become blocked without either party being aware of it. The production assistant's report does not get through, and the marketing assistant misses an opportunity to expand output to meet buoyant demand. The production assistant will simply think that demand must have been depressed, while the marketing assistant believes that technology was unfavourable anyway. Neither assistant is surprised by the outcome. The situation can persist over several periods until the owner becomes suspicious that the run of apparently unfavourable states is unusually long in the light of his subjective beliefs.

If the channel is blocked in the other direction then the mistake will be corrected after one period. If the marketing assistant's reply to a favourable report on technology does not get through then high output will be produced using the labour-intensive technique. The owner will realize that a mistake has been made because he knows that this combination is never efficient. The blockage is therefore likely to be discovered and put right.

The restriction of communication to exceptional cases reduces the information cost associated with non-decisiveness, and so discourages the imposition of decisive rules. It actually encourages what was earlier described as a consultative approach, although what now pases for consultation is, under normal circumstances, just a set of mutually compatible beliefs based on no real communication at all. In other words, when the non-decisive rules operate under normal circumstances, everything automatically takes care of itself.

9. Investigation as a Sequential Process

So far the only kind of information cost that has been considered is the cost of tacit communication. But information that is communicated has to be discovered first. While discovery of the factor states may be a byproduct of other activities performed by the assistants, such discovery is not without its cost, as noted earlier. The question naturally arises, therefore, as to whether some decision rules economise on investigation cost more than do others (Hirshleifer and Riley, 1992).

The answer is affirmative. States only need to be investigated if the decision rule conditions on them. If the same state is decisive for both strategic variables, for example, then the other state is irrelevant to both decisions and hence does not need to be investigated at all. Thus if demand is decisive for both output and technique, say—$x_1 = x_1(s_1)$, $x_2 = x_2(s_1)$—then no information is required on technology. This situation has already been encountered in region 2 of Figure 2.

Even greater savings are available if neither state is relevant. Thus region 4 in Figure 1 corresponds to a situation where low output produced with the labour-intensive technique is always the best policy whatever the state of technology or demand. Variations in these factors cannot overturn the basic logic of this production plan because both the price premium and the capital cost premium are too low.

Intuitively this shows that factors which have little impact are irrelevant to the decision. But it is possible to go further than this, and note that even factors which are potentially relevant may not be worth investigating, because the expected gain from conditioning strategy on them may be outweighed by the cost of investigation. This illustrates the principle that was encountered earlier in the context of communication costs: that it may be useful to suppress information of potential relevance if the expected gains are lower than the costs incurred.

The expected gains vary, as noted above, with respect to the premia involved. But they also vary according to the probabilities with which relevant situations occur. If states in which the results of the investigation would make a difference to the decision are subjectively unlikely, then costly investigation will not normally be worthwhile.

Investigation, like communication, can be carried out sequentially. This allows the decision whether to make the second investigation to be conditioned on the first. This in turn raises the issue of the order in which the investigations are best carried out. This highlights another dimension to the question of procedures. It is not just a question of whether sequential procedures are advantageous, but which of various alternative procedures is the best. (Although this issue arises in the context of communication too, it is of less significance for decisiveness there).

Three main sequential investigation strategies can be identified. The first is to investigate s_1 and then decide, in the light of this, whether to investigate s_2. The second is to investigate s_2 first and then decide whether to investigate s_1 as well. The third is to do no investigation at all. The first two strategies may be termed option strategies since they defer the second decision so that it can be conditioned on the outcome of the first. For each option strategy there are four possible decision rules that can be used at the second stage. Choosing the best of these four rules is akin to the selection of a stopping rule in a search process (Lippman and McCall, 1976).

The choice between the options is equivalent to the selection of a 'where to start' rule. This is important in the present context because the information on each factor is qualitatively different from that on the other: the second observation is an observation on a second variable, rather than a second observation on the same variable, as in conventional search theory.

Solving for the optimal investigation strategy involves two main steps. The first step is to determine the optimal output and technique conditional on each possible set of information on the factors. The option approach is then invoked to solve recursively for the optimal investigation strategy, assuming that the strategies derived in the first step are always employed.

If communication costs are allowed for then the first step becomes formidably complicated because of the various strategic issues that must be addressed in each of the many different cases that can arise. Communication costs are therefore ignored in the analysis of investigation. This restriction means that where decisiveness is imposed as part of the solution it will be purely on account of investigation costs, and not communication costs as well.

Consider now the numerical example given in Table 3. Recall that in this example demand is naturally decisive for output, so that the only reason for investigating technology is that the choice of technique may be improved as a result. Investigation costs are therefore likely to impact mainly on the production assistant, who will be denied the opportunity to investigate technology. He may simply have to take his cue from the marketing assistant, whose information on demand is treated as decisive for technique even though naturally this is not the case. Formal analysis does indeed confirm this intuition, but indicates some other interesting possibilities too.

To fix ideas, suppose that the cost of investigation in either state is 0.1. The top half of Figure 5 illustrates the determination of the best variant of option 1. The left hand quadrant indicates what happens if the initial investigation reveals depressed demand $s_1 = 0$. Only two operational strategies merit consideration: the low-output labour-intensive strategy indicated by the line AB and the low-output capital-intensive strategy indicated by the line CD. These dominate both the

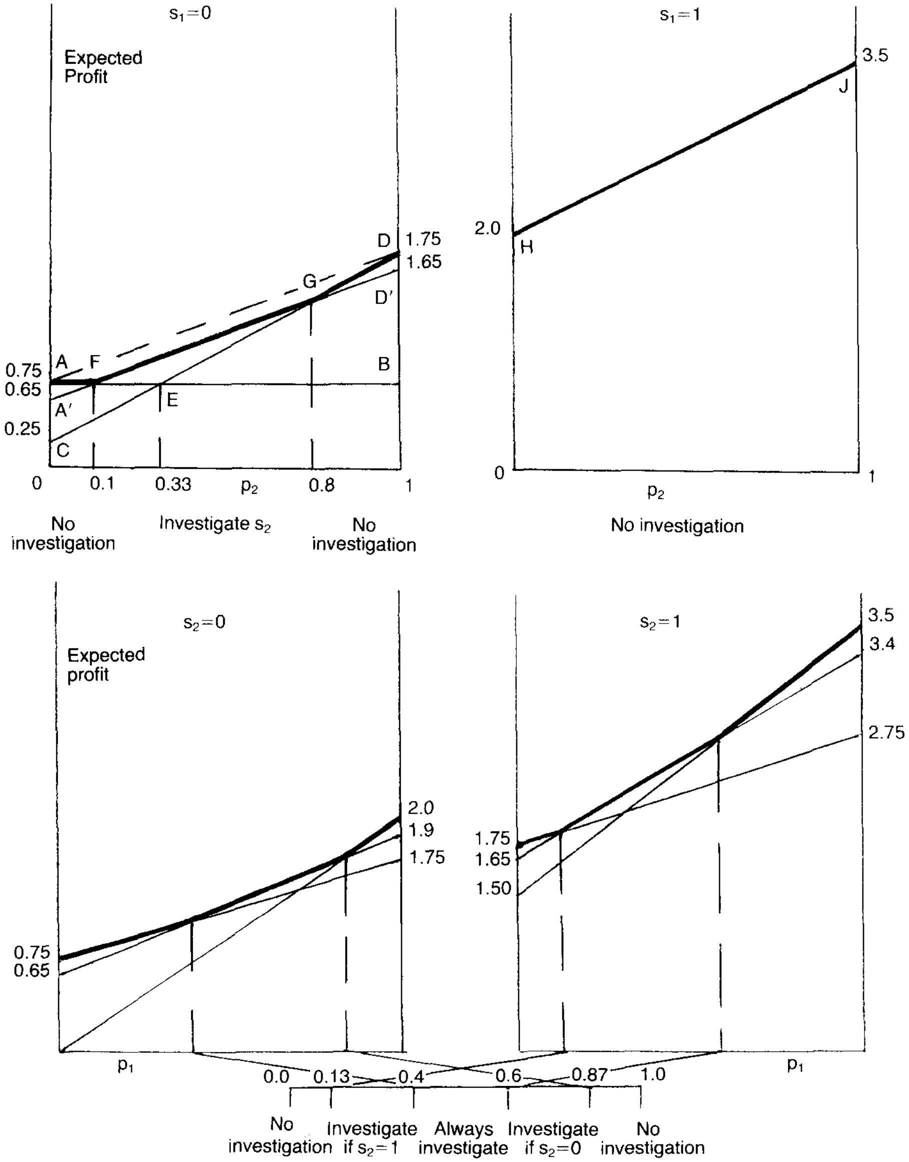

Figure 5. (Top) Option 1: investigate s_1 first. (Bottom) Option 2: investigate s_2 first.

high-output strategies when demand is depressed. The labour-intensive strategy offers the highest expected value when technology is likely to be unfavourable ($p_2 \leq 0.33$); otherwise the capital-intensive strategy appears best.

The advantage of investigation is that it is unnecessary to gamble on which state of technology prevails. Prior to investigation it is known that the best technique can be chosen in the light of the information obtained. In the absence of investigation costs this would generate the valuation line AD, but allowing for these costs shifts the line down to A′D′. This generates the valuation frontier AFGD, with kinks at F and G corresponding to critical probabilities 0.1 and 0.8. For values of p_2 within this range further investigation is the correct response to $s_1 = 0$; outside this range it is better to take a gamble instead.

The top right-hand quadrant shows that if buoyant demand is revealed, $s_1 = 1$, then one operational strategy dominates all others, namely high output produced with the capital-intensive technique. Attaching a weight $1 - p_1$ to the valuation AFGD and a weight p_1 to the valuation HJ determines the valuation of option 1.

The valuation of option 2 is derived by a similar method in the lower half of the figure. It can be seen that the strategic response to an initial investigation of technology is somewhat more complicated: this is because technology, unlike demand, is not conditionally decisive. For values of p_1 between 0.4 and 0.6 it pays to investigate demand, s_1, whatever the state of technology, s_2, turns out to be. For $0.13 \leq p_1 < 0.4$ it pays to investigate demand only if technology is favourable ($s_2 = 1$), while for $0.6 < p_1 \leq 0.87$ it pays to do so only if technology is unfavourable instead ($s_2 = 0$). Weighting the two valuation functions (indicated by the emboldened lines), according to the value of p_2 gives the expected value of option 2.

The valuation of the 'no investigation' strategy is summarised in Figure 6. It portrays the unit square of $p_1 - p_2$ space, as shown earlier in Figure 4. The figure is divided into three segments, revealing that only three of the four operational

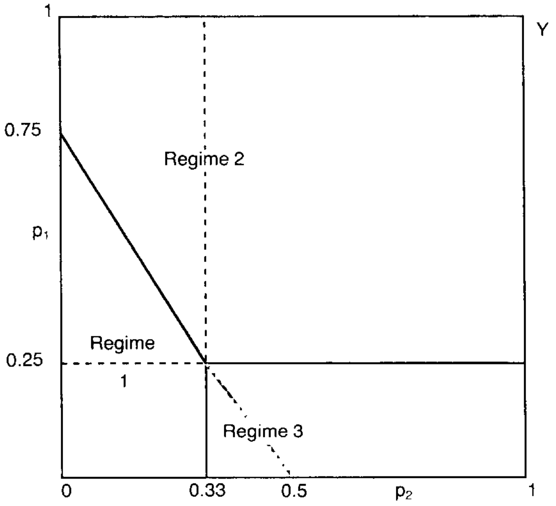

Figure 6. Operational strategies when there is no investigation. *Note:* In regime 1: $v = 0.25 + p_1$; $x_1 = 0$, $x_2 = 0$. In regime 2: $v = -0.5 + 2p_1 + 1.5p_2$; $x_1 = 1$, $x_2 = 1$. In regime 3: $v = -0.25 + p_1 + 1.5p_2$; $x_1 = 0$, $x_2 = 1$.

strategies is ever viable in an uncertain situation. The combination of high output and labour-intensive technique is always dominated by one of the others—but which one depends on the subjective probabilities involved.

Comparing the valuations of the three investigation strategies indicates how the optimal strategy varies according to these subjective probabilities. Figure 7 divides the unit square into five regions. Option 1 is preferred in the central horizontally-shaded region ABCDEFML where both technology and demand are uncertain. Option 2 is preferred in the lower-middle vertically-shaded region FGHJKM where technology is uncertain but demand is likely to be low. This is because technology affects the choice of technique only when demand is low.

Option 3 is preferred in the three unshaded regions on the periphery of the figure. Each of these regions corresponds to a different operational strategy, as identified in Figure 6. The region OLKJH, which applies when the owner is pessimistic about both demand and technology, corresponds to the low-output labour-intensive strategy associated with regime 1 in Figure 6. The much larger region ABCDYX, which applies when the owner is optimistic about demand,

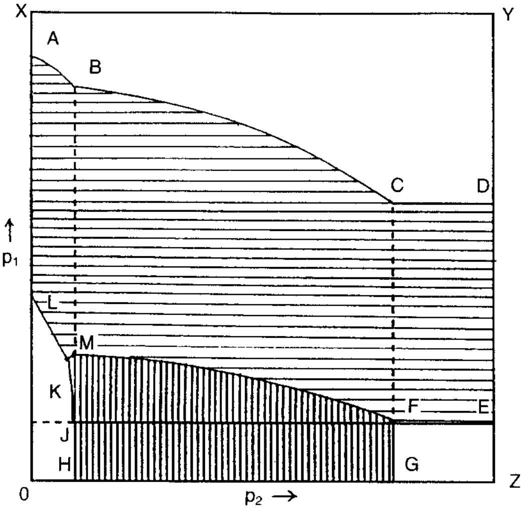

Figure 7. Choice of investigation strategy. *Key:* ▤ investigates s_1 first; ▥ investigates s_2 first; □ no investigation.

Note:	Equations of segments	Coordinates
AB	$p_1 = (0.65 - 1.5p_2)/(0.75 - 1.5p_2)$	A (0.87, 0.0)
BC	$p_1 = (0.55 - 0.5p_2)/(0.65 - 0.5p_2)$	B (0.83, 0.1)
CD	$p_1 = 0.6$	C (0.6, 0.8)
EF	$p_1 = 0.13$	D (0.6, 1.0)
FG	$p_2 = 0.8$	E (0.13, 1.0)
HJ	$p_2 = 0.1$	F (0.13, 0.8)
JK	$p_1 = (0.1 - 0.9p_2)/0.75p_2$	G (0.0, 0.8)
KL	$p_1 = 0.1/(0.25 + 1.5p_2)$	H (0.0, 0.1)
KM	$p_1 = 0.9p_2/(0.25 + 0.75p_2)$	J (0.13, 0.1)
MF	$p_1 = 0.1(1 - p_2)/(0.35 - 0.25p_2)$	K (0.27, 0.09)
		L (0.4, 0.0)
		M (0.28, 0.1)

corresponds to the high-output capital-intensive strategy associated with regime 2. Finally the corner region EFGZ, which combines technological optimism with market pessimism, corresponds to the low output capital-intensive strategy of regime 3.

It can be seen that option 3 is chosen whenever the owner is reasonably certain about what both demand and technology are going to be. This illustrates the important result that a confident owner, who believes that he knows what investigations will reveal, will tend not to investigate at all. He will simply tell the production and marketing assistants what output and technique should be, based on his estimate of the most likely case. The actual choices he imposes will vary according to his beliefs. In each case the standing orders he imposes on his two assistants will not change for as long as his beliefs remain the same.

The fact that demand is decisive in this example is illustrated by the way in which option 1, which investigates demand first, is more prevalent than option 2. It is also reflected in the way that option 3 borders the top of the figure continuously, so that optimism about demand is sufficient of itself to determine operational strategy. The other three sides are bordered discontinuously, indicating that beliefs about technology (or pessimism about demand, for that matter) do not have the same effect.

The rules that are applied to implement option 1 and option 2 vary according to the owner's subjective beliefs. Option 1 has three different versions which may be efficient, while option 2 has only two. Unlike option 3, these differences relate not merely to the implementation of the operational strategies, but to the implementation of the second stages of investigation too. Moreover, the regions that correspond to different forms of implementation are not disjoint, as with option 3, but are adjacent to one another instead.

In the central region BCFM, where technology as well as demand is uncertain, buoyant demand discovered using option 1 is taken as decisive for a high-output capital-intensive strategy. Technology is investigated only if demand turns out to be depressed. Technology guides the choice of technique, but output is set low in any case.

In the region ABML, however, where technology is likely to be unfavourable, the investigation of technology is eliminated altogether. Demand is treated as decisive for both output and choice of technique. The production assistant is no longer allowed to consult his own knowledge before choosing the technique. He is simply told that when output is low production is to be labour-intensive, and when it is high it is to be capital-intensive instead. The marketing assistant effectively runs the whole show. This case corresponds to the intuition invoked at the outset. It is interesting to note, though, that it is not quite such a common situation as intuition might suggest.

In the region CDEF the production assistant is again relegated to a minor role, but for a different reason. Because the owner is now so optimistic about technology he favours the use of the capital-intensive technique all the time. The production assistant therefore works under a standing order from the owner, and the decisions of the marketing assistant have no effect upon him.

Similar issues arise in respect of option 2. In the region FGHJ the owner is so convinced that demand will be low that he does not consider it worth investigating whatever the state of technology turns out to be. This time it is the marketing assistant, and not the production assistant, that operates under a standing order: an order to set a low level of output whatever the production assistant decides.

In the region FJKM, however, where the owner is less sure about demand, a more consultative approach prevails. Demand is investigated if technology is good, and the marketing assistant determines the output that will be produced by the capital-intensive technique that the production assistant has selected. If technology turns out to be poor, however, then the marketing assistant has to set a low level of output without first investigating demand.

10. Discrete Versus Continuous Models

In the discussion of decisiveness in Section 5 it was noted that unless the price premium and the capital cost premium exceeded certain threshold values changes in the environmental factors would not affect the firm's optimal strategy. The alert reader will have noted that this result stems directly from the discrete rather than continuous formulation of the problem. The question then arises as to whether any of the other results also depend crucially on the discrete nature of the model.

In particular, is decisiveness a property which is peculiar to discrete models and not found in continuous ones? If it were, it would suggest that the whole of the preceding analysis might be 'built upon sand'. The answer is, in fact, quite reassuring, though not as straightforward as it might be.

It is certainly possible to formulate continuous models which reveal the sort of decisiveness that generates hierarchy. Consider, for example, a firm with a family of U-shaped average cost curves, each corresponding to a different amount of capital, which generates a horizontal minimum average cost envelope. Each period capital productivity is subject to a shock whose effect on minimum average cost is exactly neutralized by a compensating change in capital good price. The price of the capital good varies in proportion to its productivity, so that when productivity falls capital input must rise to maintain output, but expenditure on capital remains unchanged. With a downward-sloping demand curve and constant-cost production, the demand parameter is decisive for output. Once output is set, the capital input is calculated with reference to the productivity parameter. Thus choice of production technique is governed by output and the state of technology. This resembles the pattern of decisiveness exemplified by Table 3.

Most conventional models of the firm, though, do not exhibit this kind of decisiveness. These models are designed for different purposes, and these purposes call for simplifying assumptions which rule decisiveness out. It is easy to show, for example, that any model of the firm in which profit is a quadratic function of output and capital input will not normally permit any of the coefficients of this quadratic function to have a decisive effect on the firm's optimal strategy. In general all the coefficients will influence each aspect of the strategy. If the problem decomposes, so that one coefficient influences just one variable, then it is because the solution separates out symmetrically, so that a hierarchical structure between different decision-makers will not appear. Only a parameter which is a function of several of these coefficients can play a decisive role of this kind. With such parameters it may indeed be possible to solve for the optimal strategy recursively: the first parameter determines the first aspect of strategy, and the first parameter and the second parameter together determine the second aspect of the strategy. But there is no reason to suppose that, in general, the factors that impinge upon the firm do so through parameters of this particular kind.

On the whole, it can be said that asymmetric decisiveness does not occur naturally in continuous models. There are exceptions, but these require very special conditions to prevail. This suggests that decisiveness is likely to emerge in continuous models only when it is imposed by the owner in order to conserve investigation or communication costs. But the construction and solution of continuous models of this kind is technically more demanding than the modelling exercises carried out above. Thus even if discrete models tend to overstate the amount of natural decisiveness, they represent a feasible solution to what may prove, in some instances, to be an otherwise intractable problem.

11. Wider Implications of the Theory

The analysis of hierarchy has some wider implications for the nature of the firm. For instance, it has been argued that the incompleteness of the employment contract is the key to understanding the nature of the firm (Williamson, 1985). But from the standpoint of hierarchy theory contracts are nowhere near as incomplete as they often seem. Each of the tasks encountered by workers in a hierarchy is, in principle, well-defined. Investigation is a well-defined activity, even though its outcome is uncertain. Rule-following is a well-defined activity provided that the rule is clearly specified. Order-giving and order-receiving are also well-defined, given that there is a finite set of possible orders that can be given or received and that order-giving is driven by rules. Any member joining an organization can, in principle, associate a subjective probability with giving or receiving any order, once his subjective probabilities relating to the different environment states are formed.

So far as a recruit to a hierarchy is concerned, therefore, the only unquantifiable uncertainty stems from the fact that the employer may suppress some relevant information about how the hierarchy works, or the kind of environment the firm is in. But this is in principle no different from someone purchasing a product which may have a hidden quality defect. In both cases some relevant information is unavailable, either because it is too costly to communicate (because it is tacit, for example) or because it is strategically withheld.

Hierarchy theory suggests that it is not so much the incompleteness of the contract that is significant for the firm, but the strategic importance of locking in the employee. The reason is that assimilation into a hierarchy incurs significant set-up costs for the employer which cannot be wholly passed on to the employee. In particular, rules have to be learnt, and so too does the specialised vocabulary in which the most commonly used orders are expressed. These are the firm-specific skills of the information-processor. In some cases they may be more costly to acquire than the manual skills of the craftsman-worker. The significance of the firm, from this point of view, is that its rules and procedures transcend the working habits of individual employees, and require the employees to conform to the institutional requirements of the firm, rather than the other way round.

Links exist with other bodies of theory, too. Note, for example, that what drives the formation of hierarchy is the need to synthesize information from different sources. An important distinction here is between sources of information on demand and sources of information on supply. These two sources need to be integrated whenever the firm is intermediating between buyers on the one hand

and sellers on the other. This is the classic role of the entrepreneurial firm (Casson, 1982). The entrepreneur recognizes an opportunity for intermediation and sets up a hierarchy to carry out this function on a continuing basis. The costs sunk in creating the hierarchy may act as a deterrent to potential entrants. The profitability of the enterprise will depend not only on the effectiveness of the deterrence, though, but also upon the efficiency with which the hierarchy is designed. This shows that great entrepreneurs need to be great organizers too.

The model shows that optimal hierarchical design is governed by the probabilities that environmental factors will be in particular states in any given period. Insufficient information is usually available to ground these probabilities in relative frequencies. They are, therefore, essentially subjective beliefs. But while individual entrepreneurs may differ in their beliefs about opportunities for market intermediation, they may well share beliefs about some of the unobservable fundamentals of the environment. This is particularly true of entrepreneurs who belong to the same social group. Thus while entrepreneurs in a given group may compete to identify opportunities for intermediation, they may well set up similar organizational structures to exploit them. The collective subjectivity engendered by the social group (Casson, 1991) may thus explain why hierarchical structures are often very similar within groups but quite different between them.

The culture of a group can influence not only individuals' subjective beliefs but also the non-pecuniary benefits that they derive from participation in corporate life. It has already been observed that hierarchies are infused with asymmetric relations involving an unequal distribution of power. Participation in these relations has an emotional dimension in which rewards and penalities can be influenced by the culture of the group. In some groups a social elite may expect to cccupy dominant roles, and the rest of the group may be content to fill the subordinate roles. This is characteristic of static 'organic' societies in which members feel an over-riding obligation to play their appointed role—a role that may even be ascribed at birth. By contrast, dynamic individualistic societies may engender a need for achievement (McClelland, 1967) which creates a sense of frustration, and even humiliation, in those required to play subordinate roles. It is this attitude that was alluded to at the outset as responsible for negative contemporary attitudes to hierarchy.

But since hierarchies have a useful role in economizing on information costs, it is worth considering whether it is the hierarchy that is in the wrong, or the negative individualistic attitude to it described above. If negative attitudes have to be compensated for by higher pay then hierarchies will become more costly to operate, and those activities which are best suited to hierarchical organization may suffer as a result. This may go some way towards explaining the productivity crisis which seems to have afflicted Western firms in a number of mass-production industries. The implications could be even wider, though, for hierarchy also seems well adapted to military activity, so that highly individualistic societies may have difficulty maintaining the hierarchical structures that they need to defend themselves effectively.

The preceding observations assume that in an individualistic society the satisfaction that superiors derive from dominating others is not so great as the dissatisfaction that subordinates feel in being dominated by others. A very aggressive individualistic society might well confer such great satisfaction on the superiors that they were willing and able to compensate their subordinates for their lower status. This would create the market in status described by Frank

(1985), in which pay differentials are narrowed within the organization to institutionalize compensation of this kind. The more individualistic the society, the greater will be the emotional responses that need to be compensated, and the more compressed the organizational reward structure will become.

If an individualistic society does not confer great emotional rewards upon superiors, however, then they will be unwilling to transfer sufficient income to buy off their subordinate's discontent. The key to the survival of the hierarchy in an individualistic society would therefore seem to depend on the emotional rewards available to superiors.

It is interesting to examine some of the recent literature on business strategy from this perspective. On the one hand, the concept of competitive strategy (Porter, 1985) is clearly hierarchical in the sense that it identifies a few key factors which it claims are decisive, and calls upon chief executives to implement strategies based on investigations of these factors, leaving subordinates to work out the tactical details. On the other hand, the fashionable concepts of empowerment and networking (Peters and Waterman, 1982) imply a much more consultative style of management in which there is less emphasis on order-giving and more emphasis on tacit communication. This is incompatible with the view that the chief executive can enunciate a strategy without first consulting extensively with other members of the organization.

The analysis in this paper suggests that the degree of empowerment should reflect the nature of the environment in which the firm has to intermediate. On the one hand, there are good reasons to believe that recent changes in the environment of business now favour a more consultative approach. The greater integration of the world economy, in terms of trade, finance and investment, means that each enterprise is exposed to a much wider range of volatile factors than before. It is therefore less likely than it used to be that just one or two factors are decisive for strategy. Furthermore improvements in business travel now make frequent facet-to-facet meetings between managers easier to arrange. This reduces the cost of consultation relative to the cost of order-giving and so makes conventional hierarchy less attractive.

On the other hand, the theory also suggests that there are dangers in pursuing the consultative approach too far. If Western culture has, indeed, created a situation where subordination breeds resentment but superiority offers few rewards then a more consultative approach will simply mitigate the worst effects and, in doing so, may postpone attempts to tackle the underlying problem. What is more, with new information technology the communication of orders has become much cheaper and, thanks to programmable computers, the implementation of rules has become cheaper and more reliable as well. Hierarchy therefore has a promising future if it is taken out of human hands. This suggests that while consultative management may indeed be the model for high-level decisions, the elimination of middle and lower management by hierarchical computer systems may be the model for lower-level decisions. Productivity losses attributable to resentful subordinates may only hasten this process of computerization.

In conclusion, there seems little doubt that hierarchy as a principle is fundamentally sound. Hierarchy is widely used as an organizing principle in nature and, as noted above, it is well adapted to implementation through computer systems. If it does not work well in human society then that, it may be suggested, is a weakness in human society rather than a weakness in the concept of hierarchy itself. Hierarchy is not incompatible with entrepreneurship—indeed many hierarchies have been established on entrepreneurial initiative. It is really only a hierarchy that holds a system-wide monopoly that needs to be feared—which is probably why many critics of hierarchy choose state bureaucracy as their target, rather than private firms. While improvements in tacit communication may well reduce dependence on hierarchy, therefore, they are unlikely to eliminate it altogether.

12. Summary and Conclusion

This paper has shown that the exploitation of decisiveness holds the key to the emergence of hierarchy. An item of information is decisive for a strategic decision when it is both necessary and sufficient for optimizing the decision, given that its tactical implementation will be optimal in the light of the other information that is available. This other information is dispersed around the organization. Decisiveness implies that it is not necessary to consult other members who hold this information before taking the decision. When decisiveness is asymmetric then these other members need to know this decision before they can take their own, however. It is the decision and not the information itself which is communicated to them because the information is tacit, and hence costly to communicate, whereas the decision is not.

Decisiveness is sometimes a natural property of an economic problem but more often than not it is simply imposed in order to economize on information costs. Information costs include both communication costs and investigation costs. Both can be minimized by a sequential approach. In addition, communication costs can be reduced by reporting only exceptional information. Economies in communication can generate a hierarchical structure independently of decisiveness, but with the difference that superiors implicitly consult with their subordinates before making decisions. This process of consultation is highly 'procedural', however.

There are two distinct dimensions to a hierarchy. One is concerned with the way that the owner of the firm hands down rules to his employees and the other with the way that the employees give orders to one another. There is always a risk that the employees may not follow the rules, or may ignore the orders they are given. This creates a logically separate hierarchy of supervision which is concerned with processing information on employee behaviour rather than with processing information on the market environment itself. It is important not to confuse the two hierarchies, as many writers on organization seem to have done.

Optimal hierarchical structure involves a trade-off between the quality of decision-making and the information costs incurred. (In a more sophisticated analysis agency costs could be introduced as well). This trade-off is governed by the nature of the economic problem that the firm is trying to tackle. This paper has focussed on the problem of entrepreneurial intermediation, rather than on production management, which is the more usual focus of hierarchical analysis.

Intermediation synthesizes information from the two sides of the market. A key question for the owner is whether demand factors are more volatile than supply factors. If he believes they are, then the firm is likely to be marketing-led, with marketing personnel acting decisively and relegating production personnel to a subordinate role. The theory therefore not only determines how far the firm will have a hierarchical structure; it also predicts which personnel in the firm will exert most hierarchical power.

In a highly integrated economic system, the number of factors relevant to the firm's decisions is likely to be large. This will reduce the scope for decisiveness and encourage a more consultative management style. If the high level of integration is a consequence of entrepreneurial attitudes and individualistic values, then cultural aversion to subordination will reinforce the switch away from hierarchical organization. A fall in the cost of tacit communication relative to the cost of explicit communication will also play a part in this. These factors seem to explain current movements towards flatter organizational structures. Current trends in communications technology suggest that in future high-level strategy formulation is likely to become even more consultative while lower-level tacital implementation becomes, if anything, more hierarchical and certainly more computer-based.

The restriction of the model to two managerial employees means that the kind of hierarchical structure generated is an extremely simple one. It is sufficient for analysing the relation of control out of which hierarchies are fashioned, but inadequate to explain the pyramid structure as well. A complete model of a pyramid requires a minimum of three participants. The simplest way to achieve this is to allow the owner to play an active role in management—in particular for him to act as a controlling intermediary when consultation between the employees is required. It is hoped to extend the model in this direction in a future paper.

Acknowledgements

I am grateful to Peter Buckley, Janet Casson, Peter Hart, Brian Loasby, David Oakes and Michael Waterson for their comments on an earlier draft.

References

Aoki, Masanao. "Towards an Economic Model of the Japanese Firm," *Journal of Economic Literature*, 1990, 28, 1–27.

Casson, Mark C., *The Entrepreneur: An Economic Theory*. Oxford: Basil Blackwell, 1982 (reprinted Aldershot: Gregg Revivals, 1991).

Casson, Mark C., *Economics of Business Culture: Game Theory, Transaction Costs and Economic Performance*. Oxford: Clarendon Press, 1991.

Coase, Ronald H., "The Nature of the Firm," *Economica* (New Series), 1937, 4, 386–405.

Foss, Nicholai J., "Theories of the Firm: Contractual and Competence Perspectives," *Journal of Evolutionary Economics*, 1993, 3, 127–44.

Frank, Robert H., *Choosing the Right Pond*; *Human Behaviour and the Quest for Status*. New York: Oxford University Press, 1985.

Hirshleifer, Jack and Riley, John G., *The Analysis of Uncertainty and Information*. Cambridge: Cambridge University Press, 1992.

Kanter, Rosabeth M., *When Giants Learn to Dance: Mastering the Challenge of Strategy, Management and Careers in the 1990s*. London: Simon and Schuster, 1989.

Knight, Frank H., *Risk, Uncertainty and Profit*, G. J. Stigler, ed. Chicago: University of Chicago Press, 1921.

Lippman, Steven A. and McCall, John J., "The Economics of Job Search: A Survey," *Economic Inquiry*, 1976, 14, 155–89 and 347–68.

Marglin, Stephen A., "What Do Bosses Do? The Origins and Functions of Hierarchy in Capitalist Production," *Review of Radical Political Economics*, reprinted in A. Gortz, ed., *The Division of Labour: The Labour Process and Class Struggle in Modern Capitalism*. Brighton, Sussex: Harvester Press, 1976, 13–54.

Marschak, Jacob and Radner, Roy, *Economic Theory of Teams*. New Haven, CT: Yale University Press, 1972.

McClelland, David C., *The Achieving Society*. New York: Free Press, 1967.

Milgrom, Paul R. and Roberts, John, *Economics of Organization and Management*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Miller, Gary J., *Managerial Dilemmas: The Political Economy of Hierarchy*. Cambridge: Cambridge University Press, 1992.

Nystrom, Paul C. and Starbuck, William H., *Handbook of Organizational Design*. Oxford: Oxford University Press, 1981.

Peters, Thomas J. and Waterman, Ronald H., *In Search of Excellence: Lessons from America's Best-run Companies*. New York: Harper and Row, 1982.

Polanyi, Michael, *Science, Faith and Society*. Chicago: University of Chicago Press, 1964.

Porter, M. E., *Competitive Advantage*. New York: Free Press, 1985.

Radner, Roy, "Hierarchy: The Economics of Managing," *Journal of Economic Literature*, 1992, 30, 1382–1415.

Sah, Raaj K. and Stiglitz, Joseph E., "The Architecture of Economic Systems: Hierarchies and Polyarchies," *American Economic Review*, 1986, 76, 716–27.

Simon, Herbert A., *Models of Man: Social and Rational*. New York: Wiley, 1957.

Wiggins, Steven N., "The Economics of Firms and Contracts: A Selective Survey," *Journal of Institutional and Theoretical Economics*, 1991, 147, 603–61.

Williamson, Oliver E., *Economic Institutions of Capitalism*. New York: Free Press, 1985.

Copyright © 2003 EBSCO Publishing