Organizational Diseconomies of Scale

R. Preston McAfee
Department of Economics
University of Texas at Austin
Austin, TX 78712-1173

John McMillan
Graduate School of International Relations and Pacific Studies
University of California, San Diego
La Jolla, CA 92093-0519

Private information creates a cost of operating a hierarchy, which becomes larger as the hierarchical distance between the information source and the decision maker increases. When information about a firm's capabilities is dispersed among the individuals in the firm, production is inefficient even though everyone behaves rationally. Because hierarchies need rents in order to function, a firm with a long hierarchy may not be viable in a competitive industry.

1. Introduction

What are the costs of hierarchy? How can a hierarchy be less than the sum of its parts? Williamson (1985, p. 131) asked: Why can't a large firm do everything that a collection of small firms can do and more? At the level of the economy as a whole, Lange (1938) asked the same question: "Why can't a central planner mimic the market, mainly using the price system to allocate resources but sometimes intervening to produce an outcome that, in the planner's view, improves upon the market?" We offer a model in which organizational diseconomies of scale arise when people in a hierarchy exploit the bargaining power that their private information gives them. The model rationalizes the commonsense observation that longer hierarchies generate larger distortions.

Hayek (1945) argued that the costs of hierarchy arise from the fact that knowledge is dispersed among the people in the organization. By knowledge, Hayek had in mind not just scientific and engineering knowledge, but also more mundane facts; he noted that "knowledge

For comments, we thank Rich Fullerton, Bengt Holmström, Hideshi Itoh, Paul Milgrom, Barry Naughton, Yingyi Qian, Eric Rasmusen, John Roberts, Michael Rothschild, and the referees.

Journal of Economics & Management Strategy, Volume 4, Number 3, Fall 1995, 399–426.

of people, of local conditions, and of special circumstances'' is a valuable asset. Such knowledge is pervasive. A worker on the production line might observe quality-defect problems that become apparent only on the shop floor, or notice a machine that is sometimes idle, or a surplus stock of raw materials that could be used. A middle manager might become aware of engineering problems in a new process, or of a way of reassigning workers to increase productivity. Salespeople in the field learn about demand for the firm's products. Much of the information about demand and costs that the top management needs for planning must come from below.[1] Knowledge that is valuable to an organization is acquired by people—at all levels of the organization, including the lowest—as a by-product of their day-to-day duties; often it consists of information about things that are transitory and seemingly trivial.

Why does it matter that the source of the information is separated from the decision-making responsibility? Organizational costs are multifaceted. Our model focuses on one particular source of organizational diseconomies of scale. People in organizations devote energy to influencing the organization's decisions to their advantage; this is the basis of the influence–cost theory of Milgrom (1988), and Milgrom and Roberts (1988, 1990a). We shall examine a specific form of influence cost: the strategic use people make of any special knowledge they have acquired. As Hayek said, ''practically every individual has some advantage over all others because he possesses unique information of which beneficial use might be made, but of which use can be made only if the decisions depending on it are left to him or are made with his active cooperation.'' Individual incentives, we shall argue, create a fundamental impediment to efficiency in hierarchies. Dispersed information within a hierarchy makes conflicts of interest inevitable. Information becomes distorted in our model—there is, in effect, a cost of communication—not because of limits to people's ability to transmit and receive information, but because of people's incentives to exploit any informational advantages they have. This distortion increases cumulatively as the information moves up the hierarchy, so longer hierarchies have greater informational inefficiencies.

1. The importance of such information for the running of a firm is stressed by Hayek (1945), Milgrom and Roberts (1988), and Schiff and Lewin (1968, 1970). In the view of Aoki (1988, Ch. 2), Japanese firms' ability to utilize production-floor information is one of the sources of their competitive edge. Levine and Tyson (1990) review empirical studies that find that employee participation in decision-making often improves firms' productivity, in part because it makes use of knowledge about the workplace that workers have and managers lack.

The experience of large European firms exemplifies the bargaining advantages of privately-held information that are the focus of our analysis. "People are reluctant to share their information," observed the head of a French company. "Managers in particular seem to think it gives them extra power." Resistance by middle-ranking managers, according to *The Economist*, has prevented most large European companies from establishing management information systems across their subsidiaries in the various countries in which they operate. "Those lower down the management hierarchy . . . have an interest in husbanding information—and the power that goes with it. . . . The types of information which these companies have found it most difficult to standardise and collect is that on customers, pricing, product specifications, and local personnel. . . . This is precisely the information which would be most helpful in lowering costs or responding quickly to changes in the market."[2]

The model to be developed is stylized, in that it ignores other sources of hierarchy inefficiencies, such as the limits to people's ability to process information (Geanakoplos and Milgrom, 1991), and the cost of monitoring subordinates and the resulting inadequate effort levels (Qian, 1994). The model is one-sided, in that it focuses on the costs of hierarchy, but not on the benefits (apart from some analysis of coordination gains in Section 6); in particular, technological economies of scale are ignored.

In a firm, according to our model, (a) production efficiency falls as the hierarchy lengthens; (b) production efficiency may rise or fall, depending on the form of the cost function, when the firm's output market becomes more competitive; (c) the longer the hierarchy, the smaller the marginal rate of payment with respect to output of the workers at the bottom of the hierarchy (so small firms will pay their workers piece rates, large firms will pay closer to fixed wages); (d) the more competitive the firm's output market, the more sensitive pay is to performance (so competitive firms will pay their workers piece rates, monopolists will pay closer to fixed wages); (e) the higher an individual is up the hierarchy, the more sensitive are marginal payments to performance (so bonuses will be a bigger fraction of income for executives than for production-line workers); and (f) a firm with a long hierarchy may not be viable in a competitive industry (so a large firm might respond to an increase in competition by shortening its hierarchy).

The degree of industry concentration, in our model, depends on the nature of the demand for industry's output. When the demand

2. "The Flowering of Feudalism," *The Economist*, February 27, 1993, p. 70.

curve is elastic over a broad output range, the industry will not be monopolized. This is novel, for it gives a demand-side rationale for industry concentration, which complements the conventional supply-side rationale based on economies of scale. The industry will tend to be competitive if there exist close substitutes for its output (for example, from imports); conversely, industries without close substitutes will tend to be monopolized. Thus we have a reversal in the conventional causality of perfectly competitive industries. Rents, we shall argue, are the lubricants that make it possible for a hierarchy to function. Economists usually think of an industry as being competitive because the firms in it are small. But if larger firms mean longer hierarchies, then potential rents must be present for a large firm to be viable. Thus firms are small because the industry is competitive.

2. Evidence on the Costs of Hierarchy

Since the informational rents that are the subject of this paper arise from private information, evidence on their existence or size is difficult to obtain, requiring as it does an unusual amount of inside knowledge. Such evidence must come either from an extraordinarily detailed audit of the organization's accounts, or from whistle-blowing insiders. Some evidence does exist, however, from capitalist firms, communist firms, and firms in transition economies.

A study of divisions of three large U.S. corporations by Schiff and Lewin (1968, 1970), based on interviews and examination of the accounts, finds that divisional managers built slack into their annual divisional budgets by understating expected revenues using—low price and sales estimates—and overstating costs—inflating personnel requirements, proposing unneeded projects, and failing to report the adoption of cost-lowering process improvements. On the cost side, ". . . opportunities for incorporating slack are numerous and appear to require intimate knowledge of the budget and control system." (Schiff and Lewin, 1968, p. 61). The slack was lower in years in which operating conditions were adverse than in favorable years, consistent with the model to be developed. This slack was large, amounting to an estimated 20% to 25% of the division's budgeted operating expenses. In terms of the model to be developed below, if we interpret the division as the agent and the top management as the principal, this implies that the agent's informational rents averaged one-quarter to one-third of the actual production cost. The top management apparently understood that the budget was being padded but, like the principal in our model, did not have precise enough information to be able to put a stop to it.

Misreporting was rife in the pre-reform Soviet firm, according to Berliner's 1957 study, based on interviews with expatriate former managers. One of Berliner's informants said there is "An enormous amount of falsification in all branches of production and in their accounting systems . . . everywhere there is evasion, false figures, untrue reports." (Berliner, 1957, p. 161). Enterprise managers misrepresented their firms' costs in their reports to the ministries. They exaggerated their needs for labor, materials, and equipment; failed to report improvements in techniques; concealed the productivity of new machines; understated the number of engineers on hand; and overstated the time needed for a task (Berliner, 1957, pp. 82–91). The misinformation caused the Soviet planer, like the principal in our model, to order inefficient output quantities (Berliner, 1957, p. 325). The misreporting was not unknown to the ministry/principal, but the manager/agent understood incentive compatibility: "Although the purchasing organizations sometimes make attempts to check up on the statements of requirements presented to them, they have no data for this purpose, and therefore they simply adopt the method of indiscriminate cutting, which in turn causes some enterprises to present even more greatly inflated statements of requirements." (Berliner, 1957, p. 91). The ministry/principal did not, however, appear to design incentive contracts like those in the model, but rather simply asked what the enterprise's technical possibilities were. The manager/agent was not rewarded for revealing production capacity to be large (Berliner, 1957, pp. 76–77).[3]

Consistent with the magnifying of informational rents that we shall find in our multi-tier model, misreporting within the Soviet enterprise " is not confined to one level of management but permeates the whole system. Within the enterprise each official seeks to maintain a little factor of safety unknown to his immediate superior. The consequence is a cumulative discrepancy between actual capacity and plan targets." (Berliner, 1957, p. 83; see also Litwack, 1989). The cumulative increase in misreporting did not even end at the enterprise level. The ministry officials in charge of the enterprise overstated its costs to the State Planning Commission (Berliner, 1957, pp. 249–251).

China's economic reforms provide an experiment in changes in hierarchy. Before the reforms, China's economic decision making relied on the flow of information through bureaucratic channels from production and consumption units. According to Naughton (1991),

3. The principal's inability to commit to the incentive scheme that will apply in the future exacerbates such misreporting; on contracting in the face of the ratchet effect, see Dearden, Ickes, and Samuelson (1990).

"the narrow channels connecting subordinates to superiors become clogged with pseudo-information, which is often intentionally distorted. While the system continues to report thousands of "bits' of data, the actual information content is quite limited." The reforms in the 1980s shifted the right to make output decisions from the state down to the firm's manager, thus eliminating a layer of hierarchy. Groves, Hong, McMillan, and Naughton (1994), testing the model to be developed below, find that managers responded to the grant of output autonomy by strengthening their workers' incentives (specifically, increasing the use of bonus payments), and the workers responded to the new incentives, significantly increasing their productivity.

3. Modeling the Costs of Hierarchy

To see why it matters that there is a separation between the decision maker and the holder of the information, we first review the case of a simple two-person hierarchy, consisting of a principal and an agent (manager and subordinate, or central planner and firm). We shall refer to the principal as "she" and to the agent as "he." Both principal and agent are assumed to be risk neutral. (In the multi-tier case to be developed in the next section, however, we shall assume that people in the middle of the hierarchy have limited liability.) The agent is better informed about, in Hayek's phrase, "the particular circumstances of time and place" than the principal; specifically, the agent has more precise information about the current level of production cost. (Alternatively, to be consistent with some of the examples given above, the private information could be modeled as being about the demand function.) The principal, operating under this informational handicap, decides how to remunerate the agent; in turn, this determines how much output the agent decides to produce.[4]

We represent the informational asymmetry by supposing the agent has a type (e.g., inherent productivity), denoted t, which determines the production cost. The agent knows his type; that is, the value of t. The principal perceives the agent's type as being drawn from a distribution $F(t)$, with density $f(t)$ and support [0, 1]. Let $C^0(q, t)$ be the cost to the agent of obtaining a given output q when his type is t. We assume that higher types have a lower cost and a lower marginal cost: $C_t^0(q, t) \leq 0$, $C_{qt}^0(q, t) \leq 0$ (where subscripts denote par-

4. The analysis in this section is standard—it builds on Laffont and Tirole (1986)—but it is necessary to develop this standard case fully as it is the basis of the induction argument for the multilevel hierarchy that follows in the next section, and of the analyses of organizational costs in the subsequent sections.

tial derivatives). We assume also that the marginal cost of output is positive and nondecreasing: $C^0_q(q, t) > 0$, $C^0_{qq}(q, t) \geq 0$. The (exogenously determined) revenue the principal earns from selling the output is $R^1(q)$ (monotonicity of the quantity q in type t follows from the supermodularity of the agent's profit function, a consequence of $C^0_{qt} \leq 0$; see Milgrom and Roberts, 1990b).

The principal, as the Stackelberg leader, specifies the agent's incentive scheme: that is, a nonlinear function stating how the size of the payment he receives will depend on the output he delivers. Denote this (endogenous) payment function by $R^0(q)$. The agent, having accepted the contract, then produces the output and is paid according to the prespecified payment function.[5]

The fundamental result is that the agent earns rents from his private information: knowledge conveys bargaining power. Because of her informational handicap, the principal cannot design an incentive scheme that extracts all of the rents. The principal, in getting the output, must not only cover the production cost actually incurred by the agent, but also offer some profit to the agent—in effect a bribe to prevent the agent from acting as though his costs are higher than they really are.

Let $\pi^0(q, t)$ be the profit earned by the agent if his type is t and he produces output q, given the payment function designed by the principal. Thus

$$\pi^0(q, t) = R^0(q) - C^0(q, t). \tag{1}$$

Given the incentive scheme imposed by the principal, R^0, the agent will choose his output according to the function $q^*(t)$, maximizing π^0 for his given type t. The principal faces an individual-rationality constraint: the contract must offer the agent nonnegative rents for all possible agent types.

From the ex ante point of view of the principal, not knowing the agent's type t, the expected amount of profit left with the agent is the expected value, over the range of possible types, of $\pi^0(q^*(t), t)$, or $E\pi^0$. It is shown in the appendix that, if the agent chooses the output that is best for him, then

$$E\pi^0 = -\int_0^1 C^0_t(q^*(t), t)\, h(t)\, f(t)\, dt, \tag{2}$$

5. Melamud and Reichelstein (1989) provide conditions under which, without loss of optimality, payment can be a function of output alone. Equivalently—and consistently with the empirical examples given in the previous section—the principal could ask the agent to report his type, having announced that payment will depend on both report and output. In this case, the optimal payment function is often linear in output, although with distinct linear functions for distinct reported types: see Laffont and Tirole (1986), McAfee and McMillan (1987).

where $h(t)$ is the inverse hazard rate $[1 - F(t)]/f(t)$ (which is assumed to be nonincreasing). Thus, in an ex ante sense (which is the relevant sense for the principal's contract-design problem), it is *as if* the agent of type t earns a profit of $-h(t)C_t^0(q^*(t), t)$ (which is positive since C_t^0 is negative). In the terminology of Myerson (1991), define the *virtual* cost to be $C^1(q, t,)$, such that:

$$C^1(q, t) = C^0(q, t) - h(t)C_t^0(q, t). \tag{3}$$

Then, from eq. (2), the expected total cost—production cost plus agent's rents—as perceived, ex ante, by the principal is equal to the expected value of $C^1(q^*(t), t)$.

The principal designs the contract so as to maximize her expected net return, which is:

$$\begin{aligned} E\pi^1(q^*(t), t) &= E[R^1(q^*(t)) - C^0(q^*(t), t) - \pi^0(q^*(t), t)] \\ &= E[R^1(q^*(t)) - C^1(q^*(t), t)]. \end{aligned} \tag{4}$$

For each possible agent type t, the principal maximizes $R^1(q) - C^1(q, t)$. In other words, the principal, in designing the incentive scheme, acts as though the agent has a known type t and cost function $C^1(q, t)$. The asymmetric-information problem has, in effect, been converted into a full-information problem. It is as if the principal produces the output herself, but at a higher cost than the agent. The principal, because she bears the informational rents, induces the agent to produce less output than the (full-information) efficient level.[6]

Individuals' incentives, therefore, create a cost of operating a hierarchy. The holder of the information exploits the bargaining power the information gives him, earning rents; in anticipation of this, the principal manipulates the outcome. This game-playing adds an information cost to the production cost and reduces the efficiency of the organization.

In the revelation-principle analysis just given, the agent correctly reports his information to the principal and receives some rents as a reward. In the real-world firms discussed above, the agent (a middle-level manager or a worker) deliberately misreports his information. To re-interpret the model consistently with this misreporting, we can think of the agent as reporting not true cost but virtual cost. The size of the cost-padding, which Schiff and Lewin (1968, 1970) estimated to be 20% to 25% of the true cost, is $-h(t)C_t^0(q^*(t), t)$. The principal (the top manager) accepts the agent's report at face value and bases

6. Note that the payment from principal to agent, R^0 is, in expectation, equal to the principal's virtual cost C^1, but in any particular realization they will typically be different.

her output decision on it. The principal knows the reported costs are inflated, but also knows there is nothing she can do about it.

4. The Multi-Tier Hierarchy

Does it matter how many hands information passes through between information source and decision maker? Does the informational cost of hierarchy increase in the length of the hierarchy? We now lengthen the hierarchy in the model just developed, and show that the answer is yes.

Consider a hierarchy consisting of three people: a top principal, a middle principal, and an agent. If the extra layer added to the hierarchy had private information of its own, then it is obvious that lengthening the hierarchy would exacerbate the inefficiencies. To examine the pure effect of the length of the hierarchy, therefore, we hold constant the amount of private information. Only the agent has private information. Anything the middle principal does to transform the output is observable by the top principal, so from a modeling point of view it is as if the middle principal simply passes the output up the chain.[7]

The top principal is assumed to be unable to contract directly with the agent. We leave this unexplained—that is, we leave unexplained why the hierarchical structure exists—but presumably it is because of bounds on any individual's span of control. Supervision takes time, and the principal's time is limited. Once an organization reaches a large enough size, employing many agents, it is not feasible for the top principal to contract with the agents directly, and she must insert subprincipals between herself and the agents (compare with Geanakoplos and Milgrom, 1991; Qian, 1994). Our interest here is not in explaining the structure of the hierarchy, but in examining the prior question of the rents that arise, given the hierarchical structure.

Assume that, initially, the top principal designs and implements the contract for the middle principal; this contract is a function $R^1(q)$ specifying how much the middle principal will be paid as a function of the output she delivers to the top principal. Next, the middle principal designs and implements the contract for the agent; this contract is a

7. Melamud, Mookherjee, and Reichelstein (1989) model a three-tier hierarchy in which the middle agent as well as the bottom agent have private information. The hierarchy is shown to be less efficient than having the principal control both agents directly. Demski and Sappington (1987) model a three-tier hierarchy in which the top principal designs all of the contracts. The intermediate principal is able to gather improved information about the agent's productivity. The top principal must motivate the intermediate principal to acquire the information. Laffont (1988) and Tirole (1986) model collusion in a three-tier hierarchy.

function $R^0(q)$ specifying how much the agent will be paid as a function of the output he delivers to the middle principal.[8] The middle principal then delivers the output to the top principal.

The middle principal (who is offered a contract prior to learning the agent's type) is assumed to face a limited-liability constraint as well as an individual rationality constraint. If the middle principal were able to post an arbitrarily large bond, then the multi-tier hierarchy would reduce to a single-tier hierarchy, effectively by the top principal selling the hierarchy to the middle principal for a fixed fee. We assume instead that the middle principal has limited borrowing capacity, and is unable to pay such a fixed fee. We require in particular that the middle principal earns non-negative rents for every possible type of the agent.

As usual in principal-agent problems, we solve in reverse of chronological order. Consider first the middle principal's design of the contract for the agent. This is, clearly, exactly the same problem as solved in the last section. The middle principal faces the reward function R^1 and the virtual cost C^1, and chooses the payment scheme R^0 to maximize the expression (4) (given that the agent maximizes his profit as in eq. (1)).

Now consider the top principal's contract design. The top principal understands that the middle principal's profit is the difference between the payment received by the middle principal and the amount the middle principal pays the agent (which is the agent's production cost plus rent):

$$\pi^1(q, t) = R^1(q) - C^0(q, t) - \pi^0(q, t). \tag{5}$$

The top principal has a reward function R^2, and must pay R^1 to the middle principal. Let C^2 be the virtual cost associated with the middle principal's cost C^1, that is,

$$C^2(q, t) = C^1(q, t) - h(t)C^1_t(q, t), \tag{6}$$

or, for the multi-tier hierarchy,

$$C^k(q, t) = C^{k-1}(q, t) - h(t)C^{k-1}_t(q, t). \tag{7}$$

We assume producing nothing costs nothing, so $C^0(0, t) = 0$. We assume also $C^k_q(q, t) > 0$, $C^k_{qq}(q, t) < 0$, and $C^k_{qt}(q, t) < 0$. These assump-

8. In our notation, the superscript on a cost or a revenue function denotes the level of the hierarchy to which it applies. Thus R^i is the revenue received and C^i the cost incurred by a person at the ith level in the hierarchy (with 0 denoting the agent at the bottom). Note also that the revenue function of the top principal is exogenous, whereas lower-level revenue functions are designed by the person at the next level up.

tions are analogous to the standard assumptions imposed in the one-tier case to ensure the solution is monotonic. They are complicated assumptions since they involve kth derivatives of the primitive function C^0. One case that satisfies them is t uniform on [0, 1] and $C^0(q, t) = (z + 1 - t)c(q)$, for increasing concave c, with $c(0) = 0$, for which

$$C^k(q, t) = (z + 2^k(1 - t))c(q). \tag{8}$$

From the one-tier analysis, $R^0(q^*(0)) = C^0(q^*(0), 0)$. Thus the limited-liability condition for the first-level principal requires

$$0 \le R^1(q^*(0)) - R^0(q^*(0)) = R^1(q^*(0)) - C^0(q^*(0), 0). \tag{9}$$

In addition, $ER^0(q^*(t)) = EC^1(q^*(t), t)$. Thus the individual-rationality condition for the first-level principal requires

$$\begin{aligned} 0 &\le ER^1(q^*(t)) - EC^1(q^*(t), t) \\ &= -[(R^1(q^*(t)) - C^1(q^*(t), t))(1 - F(t))]_0^1 \\ &\quad + \int_0^1 (1 - F(t))[R^{1\prime}(q^*(t))q^{*\prime}(t) - C_q^1(q^*(t), t)q^{*\prime}(t) - C_t^1(q^*(t), t)]dt \\ &= R^1(q^*(0)) - C^1(q^*(0), 0) - \int_0^1 C_t^1(q^*(t), t)(1 - F(t))dt. \end{aligned} \tag{10}$$

In general, either eq. (9) or eq. (10) could be the binding constraint determining $R^1(q^*(0))$. If eq. (10) binds, the first-level principal earns zero profits and the solution is for the second-level principal to "sell" the agency to the first-level principal (that is, require a payment independent of output). In this case there is no extra distortion due to the extra layer of hierarchy; the extra layer of hierarchy adds no extra hierarchy costs. If, however, eq. (9) is the binding constraint, a longer hierarchy means greater inefficiencies.

A sufficient condition for eq. (9) to bind (that is for "selling the agency" not to be possible) is as follows. Let q_0 be the most output that the lowest type of agent (i.e., $t = 0$) could ever be asked to supply. If T denotes the top principal, q_0 is given by $R^{T\prime}(q_0) = C_q^0(q_0, 0)$, and is zero if $R^{T\prime}(0) < C_q^0(0, 0)$. Then, it will turn out, limited liability always binds if

$$C^{k-1}(q, 0) - C^k(q, 0) - Eh(t)C_t^k(q, t) \ge 0, \quad \text{for all } q, \quad 0 \le q \le q_0. \tag{11}$$

This is satisfied by the example eq. (8), so there are indeed hierarchy costs when the cost function takes this form. It is also satisfied if $q = 0$; that is, if zero output is ordered from the lowest type of agent. Given that the functional forms are such that eq. (11) holds, then (a (as shown in the Appendix) the total rents increase cumulatively up the hierarchy.

The agent exploits the bargaining power that comes from his private information to earn profits, as we saw in the last section. The middle principal, according to eq. (7) is effectively in a similar position when she contracts with the top principal (given that the limited-liability constraint is binding). The middle principal, via her contracting with the agent below her, in effect inherits the agent's information, and she uses this to extract profits for herself.

How quickly do informational rents increase as we move up the hierarchy? The jth-tier virtual cost depends on the jth derivatives with respect to t of $C^0(q, t)$ and $[1 - F(t)]/f(t)$, so little can be said in general (since C^{j+1} depends on C_t^j; in turn, C^j depends on C_t^{j-1}, etc.). In a tractable special case, however, the rents rise surprisingly quickly. Let F be the uniform distribution on [0, 1] and $C^0(q, t)$ take the form $(z + 1 - t)c(q)$, with $z > 0$ (thus making the higher derivatives zero). For this example, the cost effectively borne by the jth-tier principal is $[z + 2^j(1 - t)]c(q)$. Thus each layer added to the hierarchy doubles the informational rents borne by the overall principal (which are $2^j(1 - t)c(q)$). Informational rents increase exponentially in the length of the hierarchy.

For a more general model of a multi-tier hierarchy, imagine an organization with a pyramidal structure. A single principal oversees a certain number of subprincipals, each of whom supervises some sub-subprincipals, and so on down to several agents at the bottom of the hierarchy. The agents do the actual production, each delivering his output to his immediate supervisor, who in turn delivers it to her immediate supervisor, and so on until the output reaches the overall principal. Provided the top principal's revenue function is additively separable in the different agents' outputs, the foregoing analysis extends immediately to this case, with only notational complication.[9]

Thus our model corroborates the idea that the degree of hierarchical inefficiency depends upon how far up the hierarchy the decision maker is from the source of the information.

9. In the case of multiple agents when the agents' outputs are not additively separable, a principal who supervises several people must use more general incentive schemes than the nonlinear pricing schedules considered here. A principal must ask each of the people she supervises to report their types, and then make each supervisee's reward function depend on the other supervisees' reports. Optimal contracts in this class of problems are analyzed in McAfee and McMillan (1991). (It is shown there that, for this class of problems, there arise none of the difficulties often found in problems with multidimensional types.)

5. Incentives in the Hierarchy

Imagine lengthening a hierarchy, from two tiers to three. Does adding the extra layer to the hierarchy result in an extra output distortion? The marginal cost borne by the principal in the two-tier hierarchy is $C_q^2 = C_q^1 - h(t)C_{qt}^1$, and so $C_q^2 \geq C_q^1$ if and only if $C_{qt}^1 \leq 0$. But, as is noted in the Appendix, part of a sufficient condition for the first-order conditions to characterize the one-tier solution is $C_{qt}^1 \leq 0$. Given this, the top principal's marginal cost is higher in the two-tier hierarchy than in the one-tier hierarchy. With the concavity of the total-revenue function, this means that output falls as the hierarchy lengthens. Large firms—in the sense of firms with longer hierarchies—produce, ceteris paribus, less efficiently than small firms.

What contract does the principal offer the agent in the one-tier hierarchy? The principal can evoke her desired output by offering the agent a menu of contracts. Payment is a linear function of output, with a marginal remuneration rate of $R^{0\prime}(q^*(t)) = C_q^0(q^*(t), t)$, by eq. (1) (Laffont and Tirole, 1986; McAfee and McMillan, 1987). The reason is straightforward. The agent rationally equates the marginal cost he bears to his marginal rate of payment. If the principal wants to evoke the output q when the agent is of type t, she must offer a marginal payment rate equal to the corresponding marginal cost $C_q^0(q, t)$. This marginal payment rate can be interpreted as a piece rate, commission rate, or managerial incentive scheme. Similarly, in the three-tier hierarchy, the marginal payment rate for the middle principal is $C_q^1(q^*(t), t)$. But this marginal payment rate is equal to $C_q^0(q^*(t), t) - h(t)C_{qt}^0(q^*(t), t)$. Given that $C_{qt}^0 \leq 0$, this means that (because marginal costs rise) the marginal rate of payment rises as we move up the hierarchy. A supervisor's performance bonus always exceeds her supervisee's.

A reduction in the desired output reduces the marginal rate of payment to an agent (that is, $C_q^0(q^*(t), t)$). Thus the longer the hierarchy, the smaller the agent's marginal payment rate, given the agent's type. Small firms (that is, firms with short hierarchies) will tend to pay workers piece rates; in large firms, workers' payments are closer to fixed wages.

6. Large Firms vs. Small Firms

Our model gives one answer to the question posed by Williamson (1985) about the limitations to the size of firms. Imagine merging several firms. The merged firm ought to do at least as well as the independent firms, for one option for the chief executive is to order what are now divisions to behave exactly as they would have as separate firms. But the merged firm should do strictly better than this, for the chief executive is able to intervene selectively when there are clear gains from doing so. That there exist limits to the size of firms implies something is missing from this argument.

Suppose that in the merged firm, selective intervention from the top, which can achieve efficiency and profit gains, requires that an overall principal be added to control the merged firm, so creating an extra layer of hierarchy. Then, by our argument above, the merged firm's costs of operation incorporate an informational rent associated with the extra level of hierarchy.

To compare market coordination with coordination within a firm, let us examine more carefully an experiment of the sort proposed by Williamson. Imagine an industry consisting of n separate firms. Each of these firms consist of a single entrepreneur/worker. (This can be interpreted as a reduced-form representation of a hierarchical structure.) Each firm has private information about its own type, which determines its costs of production; and it perceives its rivals' types as being independent draws from a distribution F. The firms meet each other in the product market in asymmetric-information Cournot quantity competition. We shall compare this market with the situation after the n firms have been merged and now form the n divisions of a monopolistic firm.[10] Suppose that the problems of bargaining with private information among the n firms mean that the n independent firms could not simply form a partnership; instead, the merged firm must be controlled by a single overall principal, adding a level of hierarchy. A three-way trade-off determines how the merged firm performs in comparison with the independent firms.

Two effects work to produce gains from selective intervention. One, which could be labeled a price-coordination effect, is the standard monopoly effect: by choosing the total quantity to be supplied, the monopolist is able to extract more rents from the buyers of the industry's output than are the independent firms. The second effect can be labeled the output-coordination gain from centralization. Cournot competition creates a technical inefficiency when the firms' costs differ: firms with high costs produce too much output, and firms with low costs produce too little (holding total quantity constant). This inefficiency of competition can be ameliorated in the merged firm. The principal of the merged firm can direct the low-cost divisions to

10. To permit coordination of the workers, the merged firm uses contracts as analyzed in McAfee and McMillan (1991); see footnote 9.

produce more and the high-cost divisions less. But the inefficiencies of hierarchy mean that output-coordination gains can only partially be achieved. Full technical efficiency requires that outputs be allocated so that marginal production costs be equated across the divisions. But recall from Section 3 that the costs that the principal bears are not just production costs. Rather, the principal bears production cost plus informational rents, so that what she equates across the divisions are these marginal virtual costs. Except in the measure-zero case in which marginal information costs are the same for all divisions, the principal does not induce a fully efficient allocation of production to the divisions.[11] Working in the opposite direction to these two coordination effects is the information-cost effect derived in Section 3, which tends to make the monopoly produce less efficiently than the independent firms. This inefficiency of hierarchy tends to make the competing firms more profitable than the monopoly.

To examine this trade-off further, consider an example. There are n producing units, which we shall alternately view as independent firms and divisions of a monopoly. The monopoly is controlled by a single principal; the monopolized industry has one more layer of hierarchy than the competitive industry. The cost function is $zq + (1 - t)q^2$, where t is distributed uniformly on [0, 1]. The demand curve is linear: price is $a - bQ$, where Q is total output (the sum of the q_i's). Let π^c represent expected industry profits (averaged over types) when the firms compete as independent entities; π^m total profit (to principal and agents) when the industry is monopolized; and π^1 the profit earned by the principal when the industry is monopolized. (If π^m exceeds π^c, then it is feasible for the principal to organize a takeover of the n independent firms, paying them their stock-market value π^c, and still earning positive profit.) Whether or not monopoly does better than competition depends on the parameters. In particular (as is shown in Section A3 of the Appendix): (1) for n large, $\pi^m \geq \pi^1 > \pi^c$; (2) for b large, $\pi^m \geq \pi^1 > \pi^c$; (3) for b small, $\pi^c > \pi^m \geq \pi^1$. The first of these is easily explained. When there are many firms competing, the standard profit increase from monopolization outweighs the organizational costs. Results (2) and (3), showing the effect of the demand curve's slope on the organization of the industry, are more novel. If the demand curve has a very small slope, two of our three effects disappear. There is no Cournot inefficiency, and there are no profit

11. Consider the special case in which F is uniform on [0, 1] and the cost function is $C^0(q, t) = (z + 1 - t)c(q)$ for $z > 0$. The marginal cost perceived ex ante by the principal (from eq. (1)) is $(z + 2(1 - t))c'(q)$, and equating this across different divisions with different t's does not in general equate marginal production costs, $(z + 1 - t)c'(q)$, so there is an inefficiency.

gains from monopolizing the industry. All that remains are the hierarchical losses due to the asymmetric information. Thus, when the industry demand curve is relatively flat, the industry will not be monopolized.

Rents must exist for a long hierarchy to be viable. The general conclusion from this example is that the size of the potential rents in any given industry depends on the shape of the industry demand curve. Whether firms are large or small therefore depends, in general, not only on standard considerations such as the extent of returns to scale, but also on the nature of demand.

7. Competition and the Efficiency of Production

Do monopolists produce above minimum cost, causing a welfare loss beyond the thoroughly explored allocative inefficiencies? Conversely, does competition force minimum-cost production? Generations of economists have believed that competition provides the discipline needed to induce managers to make relatively efficient production decisions. Adam Smith said that monopoly is "a great enemy to good management, which can never be universally established but in consequence of that free and universal competition which forces everybody to have recourse to it for the sake of self-defense" (Smith, 1776, p. 165). Hicks (1935, p. 8) put it more pithily: "The best of all monopoly profits is a quiet life." Despite the familiarity of the idea that competition promotes efficiency and the fact that it has some empirical support (Scherer, 1980, pp. 464–466; Caves and Barton, 1990, Ch. 6), it still lacks a convincing theoretical basis.[12]

A natural measure of the efficiency with which a firm produces a given output is the ratio of the lowest possible cost of producing that output to the actual cost. From eq. (1) efficiency, so measured, equals $C^0/[C^0 - h(t)C^0_t]$, or $1/[1 + x]$, where x is the ratio of ex ante informational rent to production cost (i.e., $x = -h(t)C^0_t(q, t)/C^0(q, t)$). Thus the extent of inefficiency depends on the size of the informational rent relative to production cost. As informational rents decline relative to production costs, measured efficiency increases towards

12. In the model of Hart (1983), firms with separate owners and managers compete with owner-managed firms. With very risk-averse managers, slack is lowered by the existence of competition. Scharfstein (1988), however, shows that with less extreme risk aversion competition may increase managerial slack. Hermalin (1992) gives sufficient conditions for an increase in competition to decrease managerial slack, when managers are risk-averse and have commitment ability. Stole and Zwiebel (1995) show that the bargaining power of workers can generate internal inefficiencies in the firm, the extent of which varies with, among other things, the competitiveness of the product market.

one. At the outset, the principal decides the quantity she wants the agent to produce by equating marginal revenue to the marginal (virtual) cost the principal pays, C^1_q. Assume the principal faces increasing marginal costs, so that $C^1_{qq} > 0$. (C^1_{qq} equals $C^0_{qq} - \{[1 - F(t)]/f(t)\} C^0_{tqq}$, so this requires that C^0_{tqq} is, if positive, not too large.) Consider the experiment of a small rotation of the demand curve about a preexisting optimal point. This increases the demand elasticity and increases marginal revenue at that point. This means (given concavity of the total-revenue function) that the principal wants more produced than before. What is the effect of the increased production on efficiency? Efficiency, measured as $1/[1 + x]$, rises or falls with increases in output as x, the ratio of informational cost to production cost, falls or rises. Thus efficiency increases as output increases if and only if qC^0_q/C^0, the elasticity of cost with respect to output, exceeds qC^0_{tq}/C^0_t, the elasticity of the rate of change of cost with respect to type. (For example, with the cost function $C^0 = [z + 1 - t]q^2$, these elasticities are equal and efficiency is independent of output. With $C^0 = zq + (1 - t)q^2$, C^0_t is more elastic than C^0, and efficiency declines as output increases.) Thus our model gives only ambiguous support for the idea that increased competition generates increased efficiency.

If we define profit as in elementary textbooks to be revenue minus production cost, then in the model developed here firms do not maximize profit; rather, they maximize revenue minus cost inflated by the informational rent. Samuelson (1976, p. 508), in assessing the assumption of profit maximization, echoes Hicks (1935) on the monopolist's quiet life: "As soon as the firm becomes of any considerable size and begins to enjoy some control over price, *it can often afford to relax a little* in its maximizing activities." According to Samuelson, firms with less elastic demand operate less efficiently. We have seen that this is true in our model if the cost elasticity condition holds; that is, if C^0_t is less sensitive to output variations than is cost C^0 itself. But efficiency falls with changes in demand elasticity not because the firm relaxes in its maximizing activities, but rather because the informational constraints facing the firm's principal change with the firm's environment.

We noted in Section 5 that the optimal contract may be implemented by a menu of linear contracts. The optimal payment scheme varies with the firm's demand. Consider again the experiment of rotating the demand curve about the existing optimum. As demand becomes more elastic, the desired quantity $q^*(t)$ rises. Because the agent's marginal costs are increasing, this means that his marginal payment rate increases. Thus the more competitive the firm's output market, the more stringent are the contractual incentives offered to the agent.

8. Summary

We have shown how private information creates a cost of operating a hierarchy, which is larger than greater the hierarchical distance between the information source and the decision maker. We have examined how contractual incentives vary along the hierarchy. The longer the hierarchy, the smaller the marginal rate of payment with respect to output of the workers at the bottom of the hierarchy. The higher an individual is up a hierarchy, the more sensitive are marginal payments to performance. When information about the firm's capabilities is dispersed among the individuals in the firm, production is inefficient even though everyone behaves rationally. We have found that, when the firm's product market becomes more competitive (i.e., the elasticity of demand for its output rises) production efficiency rises or falls depending on the form of the cost function. The more competitive the firm's output market, the more sensitive is pay to performance. Because hierarchies need rents in order to function, a firm with a long hierarchy may not be viable in a competitive industry.

The model can be interpreted as defining what comprises a hierarchy. Who is the decision maker? The top principal is defined in our model as the person who designs the terms of the transaction, using information from below. Thus, if in a planned economy the right to set outputs is shifted down form the state to the enterprise, hierarchy has been reduced (in our terms), even if the state retains nominal control over the enterprise. Similarly, a corporate reorganization that pushed decision responsibilities down to lower-level managers would correspond, in our model, to a reduction in the length of the hierarchy, regardless of whether the firm's formal organization chart had changed. One way to reduce the informational costs of hierarchy is to avoid vertical integration by organizing production Japanese-style (see McMillan, 1990, 1995). Production takes place via a chain of subcontractors and, rather than sending all decisions up to the main firm at the top of the chain, each subcontractor is made responsible for its contracts with its own subcontractors.

Appendix

A1. Rents in the Simple Hierarchy

We derive eq. (3), the virtual cost for the simple principal-agent hierarchy developed in Section 3. The principal designs the way the agent is remunerated, offering to pay an amount that depends on the output the agent produces. The profit that will be left with an agent of type t who produces an output of q is given by eq. (1). The agent is free to choose the output; let $q^*(t)$ be the output that is optimal for the agent, given his type and the payment function. We ensure that the agent is indeed optimizing in his choice of $q^*(t)$ by applying the Envelope Theorem to eq. (1):

$$\frac{d\pi^0}{dt} = \frac{\partial \pi^0}{\partial t} = -C_t^0(q^*(t), t). \tag{A1}$$

The principal must choose a payment function such that the agent is willing to participate. This means that $\pi^0(q^*(t), t) \geq 0$. Furthermore, it cannot be in the principal's interest to offer any rents to the agent in the event that the agent has the lowest possible type: thus $\pi^0(q^*(0), 0) = 0$. From the principal's ex ante point of view, the expected profit left with the agent is:

$$\begin{aligned} E\pi^0 &= \int_0^1 \pi^0(q^*(t), t)f(t)\, dt \\ &= -[\pi^0(q^*(t), t)(1 - F(t))]_0^1 + \int_0^1 (1 - F(t))(d\pi^0/dt)\, dt \\ &= -\int_0^1 C_t^0(q^*(t), t)h(t)f(t)\, dt, \end{aligned} \tag{A2}$$

where $h(t) = [1 - F(t)]/f(t)$. (Here the second line uses integration by parts, and the third line uses (A1) together with $F(1) = 1$ and $\pi^0(q^*(0), 0) = 0$. Hence, in an ex ante sense, the rent accruing to the agent is $-C_t^0(q^*(t), t)h(t)$, and the virtual cost is this rent plus the production cost, as given in eq. (3).

The principal sets the agent's marginal rate of payment, if the agent is of type t, equal to $C_q^0(q^*(t), t)$ (for then the agent's marginal benefit equals marginal cost at the desired output level—McAfee and McMillan, 1987). The first-order conditions characterize the solution if and only if this marginal payment rate increases with the agent's type. A necessary condition for this is $dq^*(t)/dt \geq 0$ (McAfee and McMillan, 1987); that is, more output is evoked from higher types. The output the principal wants maximizes $R^1(q^*(t)) - C^1(q^*(t), t)$. Thus the optimal output satisfies $R^{1\prime}(q^*(t)) - C_q^1(q^*(t), t) = 0$. Totally differentiating this expression with respect to t, we get $dq^*(t)/dt = C_{qt}^1/[R^{1\prime\prime} - C_{qq}^1]$. Thus, with concavity of R^1, a sufficient condition for the first-order conditions to characterize the solution is $C_{qt}^1 < 0$ and $C_{qq}^1 \geq 0$; that is, the principal's cost function C^1 inherits the curvature properties we assumed (at the start of Section 3) for the actual cost function C^0. (Unfortunately, since C^1 depends on C_t^0, the signs of

C^1_{qt} and C^1_{qq} depend on third derivatives of C^0.) In the analysis of the multi-tier case to follow, we assume that these conditions on C^1 hold.

A2. Rents in the Multi-tier Hierarchy

We now consider the multi-tier case and derive the top principal's virtual cost, as given in eq. (7) in Section 4. We assume throughout that eq. (11) holds. The top principal writes a contract with the second-highest principal, who in turn contracts with the next principal, and so on down to the agent, who accepts his contract and produces the output. At each level, the contract specifies payment as a nonlinear function of output alone. We solve backwards, beginning with the contract between the agent and his immediate supervisor. Clearly this is exactly the same as the two-tier case just solved, and is summarized by eq. (A2). Now consider the contracting between a principal and her superior. The middle principal's profit for a given agent type t is given by eq. (5). The middle principal in effect implicitly "chooses" the output function (as discussed in the text): not directly, but indirectly, since by her choice of contract $R^0(q)$ she determines the $q^*(t)$ function that the agent will choose. As we saw in the two-tier case, the meaning of the virtual cost $C^1(q, t)$ is that the middle principal maximizes $\pi^1(q, t) = R^1(q) - C^1(q, t)$ pointwise: that is, the middle principal "chooses" her optimal q for each value of t by maximizing this expression. (As noted in the text, it is as if the middle principal is able to produce the output herself, at a cost, known to her, of $C^1(q, t)$).

Given the middle principal's optimization, the Envelope Theorem implies

$$\frac{d\pi^1}{dt} = \frac{\partial \pi^1}{\partial t} = -C^1_t(q^*(t), t). \tag{A3}$$

Define, for $k \geq 0$,

$$C^{k+1}(q, t) = C^k(q, t) - h(t)C^k_t(q, t). \tag{A4}$$

We show that, provided the j-level principals choose quantities q^* that maximize $R^j(q) - C^j(q, t)$, for all $j \leq k$, then the $(k + 1)$-level principal does so as well, thereby establishing the induction formula given in eq. (7).

We set the base of the induction at k; that is, we suppose the kth level principal's problem has been solved, with the following three properties:

$$R^j(0) = C^0(q^*(0), 0), \quad \text{for all } j < k. \tag{A5}$$

$$ER^j(q^*(t)) = EC^{j+1}(q^*(t), t) + C^0(q^*(0), 0) - C^j(q^*(0), 0), \text{for all } j < k. \tag{A6}$$

$$R^{j\prime}(q^*(t)) = C^j_q(q^*(t), t), \text{ for all } j \leq k. \tag{A7}$$

We now show that these properties induct to $k + 1$. Provided q^* is nondecreasing, we first show that eq. (11) implies individual rationality (i.e., condition (10)) for level k does not bind. Note that if eq. (10) binds, the $k + 1$ principal sets $R^{k\prime}(q) = R^{k+1\prime}(q)$, which from eq. (A7) and $C^k_{qt}(q, t) < 0$ implies q^* is nondecreasing. Expected profits for the k-level principal are:

$$\begin{aligned}
&E[R^k(q^*(t)) - R^{k-1}(q^*(t))] \\
&= ER^k(q^*(t)) - EC^k(q^*(t), t) - C^0(q^*(0), 0) + C^{k-1}(q^*(0), 0) \\
&= -[(R^k(q^*(t)) - C^k(q^*(t), t))(1 - F(t))]_0^1 \\
&\quad - \int_0^1 (1 - F(t))C^k_t(q^*(t), t)\, dt - C^0(q^*(0), 0) + C^{k-1}(q^*(0), 0) \\
&= R^k(q^*(0)) - C^0(q^*(0), 0) - C^k(q^*(0), 0) + C^{k-1}(q^*(0), 0) \\
&\quad - \int_0^1 (1 - F(t))C^k_t(q^*(t), t)\, dt \geq 0.
\end{aligned} \tag{A8}$$

(The last step follows from limited liability ($R^k(q^*(0)) - C^0(q^*(0), 0) \geq 0$) and eq. (11).) Since individual rationality eq. (10) does not bind, limited liability eq. (9) does, establishing eq. (A5) for $j < k + 1$:

$$R^k(0) = C^0(q^*(0), 0). \tag{A9}$$

Using equations (A9) and (A6), and the integration by parts in eq. (A8), we obtain

$$\begin{aligned}
ER^k(q^*(t)) &= ER^{k-1}(q^*(t)) + R^k(q^*(0)) - C^0(q^*(0), 0) - C^k(q^*(0), 0) \\
&\quad + C^{k-1}(q^*(0), 0) - \int_0^1 (1 - F(t))C^k_t(q^*(t), t)\, dt \\
&= EC^k(q^*(t), t) - C^k(q^*(0), 0) + C^0(q^*(0), 0) \\
&\quad - Eh(t)C^k_t(q^*(t), t) \\
&= EC^{k+1}(q^*(t), t) - C^k(q^*(0), 0) + C^0(q^*(0), 0).
\end{aligned} \tag{A10}$$

This establishes eq. (A6) for $j < k + 1$. Integrating eq. (A4) by parts, we have:

$$EC^{k+1}(q^*(t), t) = C^k(q^*(0), 0) + \int_0^1 C^k_q(q^*(t),t)q^{*\prime}(t)(1 - F(t))\, dt. \tag{A11}$$

From eq. (A10),

$$ER^k(q^*(t)) = C^0(q^*(0), 0) + \int_0^1 C_q^k(q^*(t), t)q^{*\prime}(t)(1 - F(t))\,dt$$
$$= EC^1(q^*(t), t) + \int_0^1 [(1 - F(t))[C_q^k(q^*(t), t)$$
$$- C_q^0(q^*(t), t)]q^{*\prime}(t)]\,dt. \qquad \text{(A12)}$$

(The first equality uses eq. (A11); the second uses eq. (A11) for $k = 0$.) Thus, the $(k + 1)$-level principal earns

$$\int_0^1 [R^{k+1}(q^*(t)) - R^k(q^*(t))]f(t)dt$$
$$= \int_0^1 [f(t)[R^{k+1}(q^*(t)) - C^1(q^*(t), t)] - (1 - F(t))[C_q^k(q^*(t), t)$$
$$- C_q^0(q^*(t), t)]q^{*\prime}(t)]\,dt.$$

Interpreting this as the integral of a function $H(q^*, q^{*\prime}, t)$ and applying the Euler equation, we have (with the arguments of C^j suppressed),

$$0 = \frac{\partial H}{\partial q} - \frac{d}{dt}\frac{\partial H}{\partial q'}$$
$$= f(t)[R^{k+1\prime}(q) - C_q^1] - (1 - F(t))[C_{qq}^k - C_{qq}^0]q'$$
$$+ \frac{d}{dt}[(1 - F(t))(C_q^k - C_q^0)]$$
$$= f(t)[R^{k+1\prime}(q) - C_q^1] - (1 - F(t))[C_{qq}^k - C_{qq}^0]q' - f(t)(C_q^k - C_q^0)$$
$$+ (1 - F(t))(C_{qq}^k - C_{qq}^0)q' + (1 - F(t))(C_{qt}^k - C_{qt}^0)$$
$$= f(t)[R^{k+1\prime}(q) - C_q^1 - C_q^k + C_q^0 + h(t)(C_{qt}^k - C_{qt}^0)]$$
$$= f(t)[R^{k+1\prime}(q) - (C_q^k - h(t)C_{qt}^k)]$$
$$= f(t)[R^{k+1\prime}(q) - C_q^{k+1}(q, t)].$$

This establishes eq. (A7) for $j = k + 1$, which completes the induction. (Note that the assumption that q is nondecreasing follows from eq. (A7) and $C_{qt}^k(q, t) < 0$, assumed in the text.)

A3. Gains and Losses From Merger

We now derive the conditions under which a monopolist's profit exceeds the sum of competitors' profits, as stated in Section 6. The coordination gains from having a single decision maker are traded off against the costs of the extra hierarchy. The price is $p(Q) = a - bQ$, cost $= zq + (1 - t)q^2$. Let $\alpha \equiv a - z$ (z, a chosen so that price never goes negative). Let $\beta \equiv b^{-1}$.

Competition

The i^{th} firm's profits are

$$\pi_i = E_i q_i(a - bQ) - zq_i - (1 - t)q_i^2$$
$$= [\alpha - b(n - 1)\mu]q_i - (b + (1 - t_i))q_i^2,$$

where $\mu = Eq_i^*(t_i)$. Thus

$$q_i^*(t_i) = \frac{1}{2}\frac{\alpha - b(n - 1)\mu}{b + 1 - t_i}$$

$$\mu = Eq_i^*(t_i) = \frac{1}{2}(\alpha - b(n - 1)\mu)E\frac{1}{b + 1 - t}$$
$$= \frac{1}{2}(\alpha - b(n - 1)\mu)\log(1 + \beta),$$

where,

$$E\frac{1}{b + 1 - t} = \int_0^1 \frac{dt}{b + 1 - t} = -\log(b + 1 - t)\Big|_0^1$$
$$= \log(b + 1) - \log(b) = \log(1 + b^{-1}) = \log(1 + \beta).$$

Thus,

$$\mu[2 + b(n - 1)\log(1 + \beta)] = \alpha\log(1 + \beta), \text{ or,}$$

$$\mu = \frac{\alpha\log(1 + \beta)}{2 + b(n - 1)\log(1 + \beta)}$$

$$E\pi_i = E\frac{1}{4}\frac{(\alpha - b(n - 1)\mu)^2}{b + 1 - t_i} = \frac{1}{4}(\alpha - b(n - 1)\mu)^2 E\frac{1}{b + 1 - t}$$
$$= \frac{1}{4}\left[\frac{2\mu}{\log(1 + \beta)}\right]^2 = \frac{\alpha^2\log(1 + \beta)}{(2 + b(n - 1)\log(1 + \beta))^2}.$$

Industry profits under competition are

$$\pi^c = nE\pi = \frac{n\alpha^2\beta^2\log(1 + \beta)}{[2\beta + (n - 1)\log(1 + \beta)^2]}$$
$$= \frac{n\alpha^2\log(1 + \beta)}{[2 + (n - 1)\frac{1}{\beta}\log(1 + \beta)^2]}$$

$$\lim_{n\to\infty} \pi^c = 0.$$

$$\lim_{\beta\to 0} \pi^c = 0.$$

$$\lim_{\beta\to\infty} \pi^c = \infty.$$

$$\lim_{\beta\to\infty} \frac{\pi^c}{\log(1+\beta)} = \frac{n\alpha^2}{4}.$$

$$\lim_{\beta\to 0} \frac{d\pi^c}{d\beta} = \frac{n\alpha^2}{(n+1)^2}.$$

Thus, for small β,

$$\pi^c \sim \frac{n\alpha^2}{(n+1)^2}\beta.$$

Monopoly

$$y \equiv \sum_{i=1}^{n} (1-t_i)^{-1}. \quad (\text{Note } y \geq n,\ Ey = \infty.)$$

The principal's profits are:

$$\pi^1 = (\alpha - b\Sigma q_i)\Sigma q_i - \Sigma 2(1-t_i)q_i^2.$$

$$0 = \frac{\partial \pi^1}{\partial q_i} = \alpha - 2bQ - 4(1-t_i)q_i.$$

$$q_i^* = \frac{\alpha - 2bQ}{4(1-t_i)}.$$

$$4Q = (\alpha - 2bQ)\Sigma(1-t_i)^{-1} = (\alpha - 2bQ)y.$$

$$Q[4+2by] = \alpha y, \quad \textit{or}, \quad Q = \frac{\alpha y}{4+2by}.$$

$$\alpha - bQ = \frac{\alpha(4+2by) - b\alpha y}{4+2by} = \alpha\frac{4+by}{4+2by}.$$

$$\alpha - 2bQ = \frac{\alpha(4+2by) - 2b\alpha y}{4+2by} = \frac{4\alpha}{4+2by}.$$

Thus,

$$q_i^* = \frac{\alpha}{(1-t_i)(4+2by)}, \quad \text{and,}$$

$$\pi^1 = (\alpha - bQ)Q - \Sigma 2(1-t_i)q_i^2$$

$$= \alpha\left[\frac{4 + by}{4 + 2by}\right]\frac{\alpha y}{4 + 2by} - 2\Sigma\frac{\alpha^2}{(1 - t_i)(4 + 2by)^2}$$

$$= \frac{\alpha^2}{(4 + 2by)^2}[y(4 + by) - 2y]$$

$$= \frac{\alpha^2}{4}\frac{y}{2 + by} = \frac{\alpha^2}{4}\frac{\beta y}{2\beta + y}.$$

As $n \to \infty$, $y \to \infty$ (since $y \geq n$)

$$\pi^1 = \frac{\alpha^2}{4}\frac{1}{b + 2/y} \to \frac{\alpha^2}{4b} \quad \text{as} \quad n \to \infty.$$

$$\pi^1 = \frac{\alpha^2}{4}\frac{\beta y}{2\beta + y} \to 0 \quad \text{as} \quad \beta \to 0.$$

$$\lim_{\beta \to 0}\frac{d}{d\beta}\pi^1 = \lim_{\beta \to 0}\frac{\alpha^2}{4}\frac{y(2\beta + y) - 2\beta y}{(2\beta + y)^2}$$

$$= \frac{\alpha^2}{4}\lim_{\beta \to 0}\frac{y^2}{(2\beta + y)^2} = \frac{\alpha^2}{4}.$$

Thus, for small β,

$\pi^1 > \pi^c$ if and only if $\frac{n}{(n + 1)^2} < \frac{1}{4}$, which is true for $n \geq 2$.

Total Monopoly Profits

If, in order to buy the competitive firms, the top principal must only pay the difference between what agents expect under competition and what they expect under the top principal, the total profits of the monopoly are the relevant comparison to π^c (i.e., is there anything left over for the top principal?). These profits are

$$\pi^m = (\alpha - bQ)Q - \Sigma(1 - t_i)q_i^2$$

$$= \alpha\frac{4 + by}{4 + 2by}\frac{\alpha y}{4 + 2by} - \Sigma(1 - t_i)\frac{\alpha^2}{(1 - t_i)^2(4 + 2by)^2}$$

$$= \frac{\alpha^2}{(4 + 2by)^2}[y(4 + by) - y]$$

$$= \frac{\alpha^2(3 + by)y}{(4 + 2by)^2} = \frac{\alpha^2}{4}\frac{3y + by^2}{4 + 4by + b^2y^2}$$

$$= \frac{\alpha^2}{4}\frac{3y^{-1} + b}{4y^{-2} + 4by^{-1} + b^2}.$$

As noted earlier,

$$\frac{n\alpha^2}{4} = \lim_{\beta\to\infty} \frac{\pi^c}{\log(1 + \beta)} = \lim_{\beta\to 0} \frac{\pi^c}{\log(1 + b^{-1})}$$

$$\lim_{b\to 0} E\pi^m = E \lim_{b\to 0} \pi^m = \frac{3}{16}\alpha^2 Ey = \infty.$$

Although y is an improper random variable (no mean), y^{-1} is not. Since $y \geq n$, $y^{-1} \in [0, 1/n]$. Moreover $Ey^{-1} > 0$, since there is a positive probability that $y \leq 2n$.

$$\lim_{b\to 0} \frac{E\pi^m}{\log(1 + b^{-1})} = E \lim_{b\to 0} \frac{\pi^m}{\log(1 + b^{-1})} = E \lim_{b\to 0} \frac{\frac{\partial}{\partial b}\pi^m}{\frac{\partial}{\partial b}\log(1 + b^{-1})}$$

$$= \frac{\alpha^2}{4} E \lim_{b\to 0} \frac{4y^{-1} + b}{(2y^{-1} + b)^3} \Big/ \frac{1}{b(b + 1)}$$

$$= \frac{\alpha^2}{4} E \lim_{b\to 0} \frac{(4y^{-1} + b)b(b + 1)}{(2y^{-1} + b)^3} = 0.$$

Thus $\lim_{b\to 0} E\pi^m/\pi^c = 0$, and thus, for small b:

$E\pi^1 \leq E\pi^m < \pi^c$.

This gives the following summary:

(i) For n large, $\pi^m \geq \pi^1 > \pi^c$.
(ii) For b large, $\pi^m \geq \pi^1 > \pi^c$.
(iii) For b small, $\pi^1 \leq \pi^m < \pi^c$.

References

Aoki, Masahiko, 1988, *Information, Incentives, and Bargaining in the Japanese Economy*, New York: Cambridge University Press.

Berliner, Joseph S., 1957, *Factory and Manager in the USSR*, Cambridge: Harvard University Press.

Caves, Richard E. and David Barton, 1990, *Efficiency in U.S. Manufacturing Industries*, Cambridge: MIT Press.

Dearden, James, Barry W. Ickes, and Larry Samuelson, 1990, "To Innovate or Not To Innovate: Incentives and Innovation in Hierarchies," *American Economic Review*, 80, 1105–1124.

Demski, Joel S. and David E.M. Sappington, 1987, "Hierarchical Regulatory Control," *Rand Journal of Economics*, 18, 369–383.

Geanakoplos, John and Paul Milgrom, 1991, "A Theory of Hierarchies Based on Limited Managerial Attention," *Journal of the Japanese and International Economies,* 5, 205–225.

Groves, Theodore, Yongmiao Hong, John McMillan, and Barry Naughton, 1994, "Autonomy and Incentives in Chinese State Enterprises," *Quarterly Journal of Economics,* 109, 183–209.

Hart, Oliver D., 1983, "The Market Mechanism as an Incentive Scheme," *Bell Journal of Economics,* 14, 366–382.

von Hayek, F.A., 1945, "The Use of Knowledge in Society," *American Economic Review,* 35, 519–530.

Hermalin, Benjamin E., 1992, "The Effects of Competitive Pressures on Executive Behavior," *Rand Journal of Economics,* 23, 350–365.

Hicks, John R., 1935, "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica,* 3, 1–20.

Laffont, Jean-Jacques, 1988, "Hidden Gaming in Hierarchies: Facts and Models," *Economic Record,* 64, 295–306.

Laffont, Jean-Jacques and Jean Tirole, 1986, "Using Cost Observation to Regulate Firms," *Journal of Political Economy,* 94, 614–641.

Lange, Oskar, 1938, "On the Economic Theory of Socialism," in B.E. Lippincott, ed., *On the Economic Theory of Socialism,* Minneapolis: University of Minnesota Press.

Levine, David I. and Laura D'Andrea Tyson, 1990, "Participation, Productivity, and the Firm's Environment," in A.S. Blinder, ed., *Paying for Productivity,* Washington, D.C.: Brookings.

Litwack, John M., 1989, "Adverse Selection and 'Vedomstvennost'," unpublished, Stanford University.

McAfee, R. Preston and John McMillan, 1987, "Competition for Agency Contracts," *Rand Journal of Economics,* 18, 296–307.

McAfee, R. Preston and John McMillan, 1991, "Optimal Contracts for Teams," *International Economic Review,* 32, 561–577.

McMillan, John, 1990, "Managing Suppliers: Incentive Systems in Japanese and U.S. Industry," *California Management Review,* 32, 38–55.

McMillan, John, 1995, "Reorganizing Vertical Supply Relationships," in Horst Siebert, ed., *Trends in Business Organization,* Kiel: Institut für Weltwirtschaft.

Melamud, Nahum, Dilip Mookherjee, and Stefan Reichelstein, 1989, "Hierarchical Decentralization of Incentive Contracts," unpublished, Stanford University.

Melamud, Nahum and Stefan Reichelstein, 1989, "Value of Communication in Agencies," *Journal of Economic Theory,* 47, 334–368.

Milgrom, Paul R., 1988, "Employment Contracts, Influence Activities, and Efficient Organization Design," *Journal of Political Economy,* 96, 42–60.

Milgrom, Paul R. and John Roberts, 1988, "An Economic Approach to Influence Activities in Organizations," *American Journal of Sociology,* Supp. 94, S154–S179.

Milgrom, Paul R. and John Roberts, 1990a, "Bargaining and Influence Costs and the Organization of Economic Activity," in J. Alt and K. Shepsle, eds., *Rational Perspectives on Political Economy,* Cambridge: Cambridge University Press.

Milgrom, Paul R. and John Roberts, 1990b, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica,* 58, 1255–1278.

Myerson, Roger B., 1991, *Game Theory,* Cambridge: Harvard University Press.

Naughton, Barry, 1991, "Hierarchy and the Bargaining Economy: Government and Enterprise in the Reform Process," in M. Lampton and K. Lieberthal, eds., *Bureaucracy, Politics, and Decision-Making in Post-Mao China,* Berkeley: University of California Press.

Qian, Yingyi, 1994, "Incentives and Loss of Control in an Optimal Hierarchy," *Review of Economic Studies,* 61, 527–544.

Samuelson, Paul A., 1976, *Economics,* 10th edn., New York: McGraw Hill.

Scharfstein, David, 1988, "Product Market Competition and Managerial Slack," *Rand Journal of Economics,* 19, 147–155.

Scherer, F.M., 1980, *Industrial Market Structure and Economic Performance,* 2nd ed., Boston: Houghton Mifflin.

Schiff, Michael and Arie Y. Lewin, 1968, "Where Traditional Budgeting Fails," *Financial Executive,* 36, 50–62.

Schiff, Michael and Arie Y. Lewin, 1970, "The Impact of People on Budgets," *Accounting Review,* 45, 259–268.

Smith, Adam, 1776, *An Enquiry into the Nature and Causes of the Wealth of Nations.* Chicago: University of Chicago Press, 1976.

Stole, Lars and Jeffrey Zwiebel, 1995, "Organizational Design and Technology Choice under Intrafirm Bargaining," *American Economic Review,* to appear.

Tirole, Jean, 1986, "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," *Journal of Law, Economics, and Organization,* 2, 181–214.

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