Hierarchies and the Organization of Knowledge in Production

Luis Garicano

University of Chicago

This paper studies how communication allows for the specialized acquisition of knowledge. It shows that a knowledge-based hierarchy is a natural way to organize the acquisition of knowledge when matching problems with those who know how to solve them is costly. In such an organization, production workers acquire knowledge about the most common or easiest problems confronted, and specialized problem solvers deal with the more exceptional or harder problems. The paper shows that the model is consistent with stylized facts in the theory of organizations and uses it to analyze the impact of changes in production and information technology on organizational design.

I. Introduction

Organizations exist, to a large extent, to solve coordination problems in the presence of specialization. As Hayek (1945, p. 520) pointed out, each individual is able to acquire knowledge about a narrow range of problems. Coordinating this disparate knowledge, deciding who learns what, and matching the problems confronted with those who can solve them are some of the most prominent issues with which economic organization must deal.

Yet, with a few recent exceptions, most previous economics literature has equated the study of organizations with the study of incentive prob-

I thank Sherwin Rosen and Kevin M. Murphy for numerous comments and suggestions. I am also indebted to Banito Arruñada, Gary S. Becker, Cristina Bellido, Patrick Bolton, Philip Bond, Robert Gertner, Tom Holmes, John Matsusaka, Derek Neal, Canice Prendergast, Michael Raith, John Roberts, Antoinette Schoar, David Soskice, Timothy Van Zandt, an anonymous referee, and numerous seminar participants for their helpful comments.

[*Journal of Political Economy,* 2000, vol. 108, no. 5]

lems. While many important insights have been obtained from this approach, a shortcoming is that hierarchical organization forms are assumed rather than obtained from the theory (see, e.g., Calvo and Wellisz 1978; Qian 1994). As a consequence, incentive-based theories have little to say on the impact of changes in information and communication technology on organizational design. For example, will cheaper communication technology make an organization taller or shorter? How will it affect the scope of production workers and managers?

An alternative approach is to set incentive issues aside and focus instead on the organization of knowledge in production. This is the approach adopted here. The starting point is the observation that production requires physical resources and knowledge about how to combine them. If communication is available, workers do not need to acquire all the knowledge necessary to produce. Instead, they may acquire only the most relevant knowledge and, when confronted with a problem they cannot solve, ask someone else. The organization must then decide who must learn what and whom each worker should ask when confronted with an unknown problem.

When classifying knowledge is cheap, figuring out where to turn when a problem solution is unknown is straightforward. Production know-how is, however, often tacit and thus is “embodied” in individuals. Knowing if someone knows the solution to a problem inevitably involves asking that person. In Section II, I show that, in this case, it is natural to organize the acquisition of knowledge as a “knowledge-based hierarchy.” In such a structure, knowledge of solutions to the most common or easiest problems is located in the production floor, whereas knowledge about more exceptional or harder problems is located in higher layers of the hierarchy. Production workers who confront problems they cannot solve refer them to the next layer of the organization, formed by specialist problem solvers. Problems are then passed on until someone can solve them or until the conditional probability of finding the solution is too low to justify continuing the search.

The key trade-off the organization confronts occurs between communication and knowledge acquisition costs. By adding layers of problem solvers, the organization increases the utilization rate of knowledge, thus economizing on knowledge acquisition, at the cost of increasing the communication required. The limited availability of time counters the increasing returns arising from fixed knowledge costs, resulting in a limited span of control of problem solvers.

The organization is characterized by the *task design,* as defined by the scope or discretionality of production workers and problem solvers and the frequency with which they actually intervene in production; and the *structure of the hierarchy,* given by the span of control of problem solvers and the number of layers in the organization. Section III studies the impact of technological changes on organizational design, including changes in the "information technology" as given by the cost of acquiring and transmitting knowledge. The model shows that decreases in the cost of both communicating and acquiring knowledge reduce the need for specialized problem solvers in the organization. These variables have, however, opposite impacts on the discretionality of problem solvers and production workers.

Cheaper acquisition of knowledge, resulting, for example, from the introduction of expert systems, increases the discretionality of each production worker and problem solver. As a consequence, production workers need to rely less often on help from specialized problem solvers. This increases the span of control of each problem solver, reduces the number of layers of problem solvers required to solve a given proportion of problems, reduces the delay needed to obtain solutions to problems, and decreases the frequency with which problem solvers intervene in the production process.

Cheaper transmission of knowledge, on the other hand, *reduces* the scope of production workers, who rely more on the (now cheaper) problem solvers. Moreover, each problem solver can solve problems for a larger number of workers, increasing his span of control.

Up to this point, the paper assumes that knowledge of higher-level workers does not need to encompass the knowledge of workers in lower levels. However, in the context of production know-how, knowledge can often be acquired only through on-the-job learning. As a consequence, the knowledge of problem solvers encompasses the knowledge of those asking them. For example, a *chef de cuisine* has usually previously been employed in all the lower-rank jobs in the kitchen. Section IV extends the results to this case. It shows that when knowledge of the higher layers must encompass the knowledge of the lower ones, the optimal organization has the same features and a structure very similar to that of the unrestricted one.

Recent work by Radner and Van Zandt[1] with a similar (non-incentive-oriented) outlook has focused on organizations as information processors. Organizations, they argue, reduce delays in information aggregation through parallel processing while increasing communication costs. The approach delivers important insights about organizations but has some unappealing features. First, it is unclear whether aggregation of information is the right metaphor for the general information processing task. At the very least, as this paper argues, this metaphor leaves

[1] Notably see Radner (1992, 1993), Radner and Van Zandt (1992), and Van Zandt (1999). Van Zandt (1998) is an excellent survey. A related literature (e.g., Crémer 1980; Geanakoplos and Milgrom 1991) considers problems of resource allocation under constraints on managerial time and suggests that hierarchical organizations increase the amount of information that can be applied to a particular decision.

out the crucial task of acquiring and transmitting knowledge and coordinating the tasks of specialized workers. Second, the kinds of organizations obtained by these papers have features that are hard to relate to real-world organizations, such as skip-level reporting (whereby a top manager often receives messages directly from far down the chain) or unbalanced networks (where managers in the same tier have a different number of subordinates). Finally, these models cannot illuminate issues of task assignment since tasks are undifferentiated.

Another paper that investigates the phenomenon of "management by exception" is Beggs (in press). This author uses queuing theory to explore the optimal allocation of workers with exogenously given skills to the different layers of a hierarchy. In contrast here, both the distribution of skills across workers and the hierarchy are endogenously derived.

More closely related to the approach of this paper is the work by Bolton and Dewatripont (1994). They build on the insight, present already in Becker and Murphy (1993), that there exists a trade-off between specialization and coordination or communication costs. In contrast to the approach I adopt here, however, they do not consider task heterogeneity directly and focus instead on a reduced form that equates specialization to higher network throughput. Moreover, the aim of the organization they study is, as in the work of Radner and Van Zandt, information processing rather than knowledge acquisition.

The paper proceeds as follows. Section II presents the model and obtains the optimal organization. Section III carries out the comparative statics analysis. Section IV extends the model to the overlapping knowledge case. Section V discusses the implications of the model. Section VI presents concluding remarks.

II. A Model of Communication and Knowledge Acquisition in Production

A. Production

Production requires physical inputs and know-how. A worker operating a machine, for instance, confronts a range of problems that must be solved in order to produce. Let $\Omega \subset R^+$ be the set of all possible problems that may be confronted and $A \subset \Omega$ be the set of problems a worker is able to solve (his "knowledge set"). Production requires that problem $Z \in \Omega$ be drawn and solved, which happens whenever $Z \in A$. Let F be the distribution of Z. I assume that this distribution is known a priori by workers, implying that workers know how "common" different solutions are. I also assume for simplicity that this distribution is continuous and nonatomic and that the corresponding density exists. To simplify notation, and without loss of generality, normalize this density so that problems are ordered from most to least "common" and the density of problems $f(Z)$ is nonincreasing. Then if the time spent in production is t_p, expected normalized output x of a single worker with constant returns to scale in production is $E[x] = t_p\int_A dF(Z)$.

Workers can learn the solutions to the problems they confront at a cost. I assume that the cost of learning an interval A of problems is proportional to the size of this interval, $\mu(A)$ (its Lebesgue measure), and call the constant per period unit learning cost c. For example, the cost of learning all problems in the interval $[0, Z]$ is cZ.

A worker in autarchy confronting such a production function learns the most frequent problems and ignores the rest. Expected net output y per unit of time is

$$E[y] = \Pr\{Z < Z_a\} - cZ_a = \int_0^{Z_a} f(\varphi)d\varphi - cZ_a. \tag{1}$$

The problem of a worker who confronts this production function is to choose the length of the interval of knowledge acquired to maximize expected output. The first-order condition of this problem is

$$f(Z_a) - c = 0. \tag{2}$$

The marginal value of acquiring knowledge is the increase in the probability that something is produced; at the optimum it equals the marginal learning cost. As figure 1 shows, the worker learns those problems that are common enough to justify their learning costs and ignores the rest.

B. Communication and Organization

Organization allows different workers to acquire different knowledge sets and communicate knowledge as required. This has two advantages: first, it allows workers to increase the utilization rate of knowledge, decreasing the per capita learning cost; second, it allows more knowledge to be acquired and used in solving problems. But it also incurs two new costs: matching the problem with the worker who knows it and communicating the answer.

I focus here on the case in which matching problems to those who know how to solve them (or "labeling" the problems) is costly. Workers then ask other workers for the solution until they find someone who

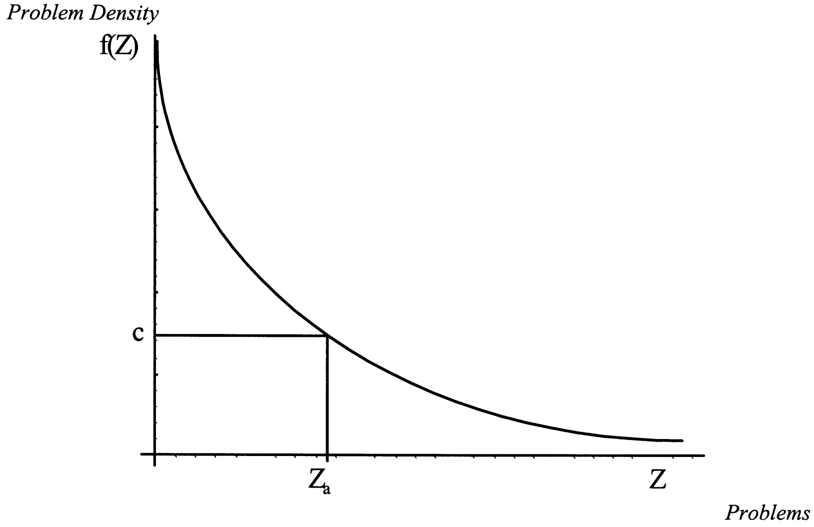

FIG. 1.—Knowledge acquired by a worker when communication is impossible

can solve it, or until they conclude that it is unlikely that anyone can solve it.[2]

The communication cost is the time spent away from production by workers communicating how to solve the problem. I make two assumptions about this cost. First, following the convention used previously in the information processing literature (Radner 1993; Bolton and Dewatripont 1994), I aggregate all the communication losses in the "receiver," in this case the worker who is being asked. Second, communication costs are incurred even when the worker asked does not know the answer, since she must figure out if she knows the answer and communicate with the worker who asked. I assume for simplicity that the cost is identical, regardless of whether the solution is known or not, and call it "helping cost" h.[3]

An *organization* is a partition of workers into L classes of size β_i (with $\sum \beta_i = 1$), such that associated with each class is (i) a knowledge set $A_i \subset \Omega$, possibly overlapping with the knowledge of workers in other classes; (ii) a list l_i of classes whom workers in i may ask for solutions,

[2] The assumption that matching problems to knowledge is hard is realistic in situations in which knowledge is hard to codify, as, e.g., when knowledge is tacit.

[3] Note that the same problem comes up again with probability zero. Thus there is no learning involved in communication.

including, in the first place, i itself; and (iii) an allocation of time to helping other classes (t_i^h) or producing (t_i^p), with $1 \geq t_i^h + t_i^p$.

I assume that the size of the organization is large enough that integer constraints can be ignored and that a law of large numbers applies to the time spent by each worker helping other workers, so that it can be dealt with as a nonstochastic variable. I also assume that everyone in the list of a particular worker may be eventually asked if necessary. This assumption seems natural since there is no point in placing someone on the list if he or she will not be asked.

The help requested by the β_k members of class k, who spend $\beta_k t_k^p$ engaged in production, to the β_i members of class i depends on the knowledge available to all classes preceding it in the list of k. Let the term $l \prec_k i$ indicate that l precedes i in the list of k. Then the time spent by workers in class i giving help to other classes is given by

$$\beta_i t_i^h = \sum_{k\,:\, i\in l_k} \beta_k t_k^p \Big[1 - F\Big(\bigcup_{l\prec_k i} A_l\Big)\Big] h \quad \text{for } i = 1, \dots, L. \tag{3}$$

Output of class i is given by the probability that at least one class in its list knows the solution to the problem confronted multiplied by the time spent by i workers in production, minus the cost of training them. Output per capita is then

$$y = \sum_{i=1}^{L} \Big[\beta_i t_i^p F\Big(\bigcup_{k\in l_i} A_k\Big) - c\beta_i \mu(A_i)\Big]. \tag{4}$$

The problem of the organization is to allocate to each class a measure of workers (β_i), knowledge (A_i), a list (l_i), and production and helping time (t_i^p, t_i^h) so as to maximize output per capita, subject to the time constraint $1 \geq t_i^h + t_i^p$ and to the organization size constraint $\sum \beta_i = 1$. The remainder of this section shows that any arbitrary original allocation of workers, knowledge, communication, and time can be improved, and thus is not optimal, unless it has the following characteristics: (1) Workers specialize either in production or in solving problems. Only one class specializes in production. (2) Knowledge acquired by different classes does not overlap. (3) Production workers learn to solve the most common problems; problem solvers learn the exceptions. Moreover, the higher up in the list of production workers a problem solver is, the more unusual the problems she is able to solve. Information in the form of solutions to problems always flows in the same direction, from the highest to the lowest level, since this minimizes communication costs. (4) The organization has a pyramidal structure, with each layer a smaller size than the previous one.

I now proceed to derive these characteristics of the organization.

Proposition 1. *Specialization.*—For any given allocation of knowledge, workers in each class specialize either in production or in the transmission of knowledge about solutions. Moreover, only one class specializes in production, and all other classes are formed by problem solvers who support workers in that class.

Proof. See the Appendix.

The intuition for this result is as follows: with knowledge per class held constant, if net output per capita of one class of workers is higher, then workers from other classes not specializing in solving problems can always be transferred to this one. This reduces the proportion of time workers remaining in the less productive class can dedicate to producing, since fewer workers must solve more problems that are asked by the larger class of productive workers. By linearity, repeating this is optimal until workers in the less productive class are able to specialize in helping the most productive class. This is true for all classes until, at the optimum, only workers in the most productive class specialize in production.

PROPOSITION 2. *Nonoverlapping knowledge.*—No solution is known by two different classes.

Proof. Knowledge of problem solvers and production workers must not overlap since problem solvers who know how to solve problems that production workers also know never use that knowledge. Knowledge of problem solvers of different classes does not overlap for a similar reason: the overlapping knowledge of the second class never gets used, and net output is higher if it is not acquired. Q.E.D.

We are now ready to obtain the key characteristic of the organization: organization by frequency or what has been called in the organizations literature "management by exception." The proof of this result relies roughly on swapping intervals of less common solutions for more common solutions between those who face a problem first (including production workers) and those who face it later so as to keep learning costs constant while decreasing the frequency of communication. As a result, net output is kept at least unchanged (organization knowledge is constant), whereas slack is created on the time constraint of problem solvers since they need to answer questions less often because those closer to production know more common problems.

PROPOSITION 3. *Organization by frequency.*—Production workers learn to solve the most common problems; problem solvers learn the exceptions. Moreover, the higher up in the list of production workers a problem solver is, the more unusual the problems she is able to solve. Information in the form of solutions to problems always flows in the same direction, from the highest to the lowest level, since this minimizes communication costs.

Proof.[4] First, to see that production workers learn to solve the most common problems, assume that they do not. Let i be the class of problem solvers who learn to solve the most common problems, so that $[0, Z_i) \subset A_i$ for some Z_i. If this class does not exist, choose any problem solver class j, and swap part or all of the knowledge set assigned to this class for an interval $[0, Z_i)$ of equal length of unlearned problems. This results in an increase in output since $\beta_w t_w^p F(\bigcup_{k \in l_w} A_k)$ has increased, leaving learning costs constant. Communication costs are reduced since all problem solvers after i answer questions less often.

Let w be the class of production workers, and let its knowledge include $[Z_w, Z'_w) \subset A_w$. Class i must belong to its list, $i \in l_w$ (otherwise it should be eliminated). The time that each i spends helping workers in w is

$$\Big[1 - F\Big(\bigcup_{l \prec_w i} A_l\Big)\Big]\frac{h\beta_w}{\beta_i}.$$

Now transfer the interval $[0, \epsilon)$ from i to w in the following way: reduce the knowledge acquired by i to $[\epsilon, Z_i)$, and swap the interval $[0, \epsilon)$ for $[Z'_w - \epsilon, Z'_w)$. Call the new knowledge set of problem solvers A'_i, $\{[\epsilon, Z_i) \cup [Z'_w - \epsilon, Z'_w)\} \subset A'_i$. Since $\mu(A_i) = \mu(A'_i)$, learning costs are constant. Do the reciprocal operation with w, so that $\{[0, \epsilon) \cup [Z_w, Z'_w - \epsilon)\} \subset A'_w$. Knowledge costs are again unchanged since $\mu(A_w) = \mu(A'_w)$. Output $\beta_w t_w^p F(\bigcup_{k \in l_w} A_k)$ is unchanged. However, slack has been created in the time constraint of all problem solvers before i in the list of w since they are now asked less often. Formally, for all $k \prec_w i$, $F(A'_w) > F(A_w)$ implies that

$$\Big[\Big(\bigcup_{j \prec_w k} A_j\Big)\Big]\frac{h\beta_w}{\beta_k} \geq \Big[1 - F\Big(\bigcup_{j \prec_w k} A'_j\Big)\Big]\frac{h\beta_w}{\beta_k}.$$

This allows some problem-solving time to be transferred to production time. This operation can be repeated until the knowledge set of production workers has the form $[0, Z_w)$.

A similar argument shows that the first place (after themselves) in the list of production workers with knowledge $[0, Z_w)$ must be occupied by problem solvers i whose knowledge includes an interval of the form $[Z_w, Z_i) \subset A_i$. If it is not, we can swap knowledge of those who know the more common problems with that of those who answer first so as to keep knowledge acquisition costs constant and reduce communication costs.

This can be generalized to levels 1, 2, and so forth. For arbitrary interval sizes, it is always better to swap knowledge in the real line so

[4] A similar argument is made in an entirely different context by Krasa and Villamil (1994) using measure theoretic tools. The present argument is sufficient given the assumptions made about the nonatomicity and continuity of the density function.

that those asked first have acquired a position closer to the origin than those asked later. The information flows from those who know the most common problems toward those who deal with the most exceptional ones. Q.E.D.

To recap, production workers always acquire knowledge about a compact set of the most common problems, and only those problems, since this minimizes communication costs. Moreover, problem solvers asked first learn relatively common problems, and those asked last deal with the most exceptional ones. This implies immediately, as will be shown in what follows, that the organization is pyramidal. But first, it helps to simplify the notation to take advantage of the fact that different classes are asked in a predictable order. Call Z_{i-1} and Z_i the endpoints of the knowledge interval of workers at layer i and $z_i = Z_i - Z_{i-1}$ the length of this interval. Then $Z_i = \sum_{j=0}^{i} z_j$.

PROPOSITION 4. *Pyramidal organization.*—An organization with multiple layers has a pyramidal structure, with each layer a smaller size than the previous one.

Proof. The proportion of workers at layer i is given by the probability that a problem has not been solved up to their level, $[1 - F(Z_{i-1})]h\beta_0 = \beta_i$, whereas the proportion of workers at layer $i+1$ is $[1 - F(Z_i)]h\beta_0 = \beta_{i+1}$. Since $Z_i = Z_{i-1} + z_i$, we know that $\beta_{i+1} < \beta_i$. In words, the higher the layer, the more exceptional the problems that are dealt with, and the smaller the proportion of problem solvers required to solve them. Q.E.D.

I have been able to characterize the solutions to the problem presented in equation (4). Figure 2 presents the general solution of the problem: L layers of problem solvers learn to solve an interval of problems $[0, Z_0)$, $[Z_0, Z_1)$, $\ldots$, $[Z_{L-2}, Z_{L-1})$. Workers in the first layer, with knowledge $[0, Z_0)$, specialize in production. The rest specialize in solving problems that production workers cannot solve. Layer 1, with knowledge $[Z_0, Z_1)$, is asked first. If those in that layer do not know the solution, layer 2 is asked, and so on.

Output can then be written as

$$y = F\left(\sum_{i=0}^{L} z_i\right)\beta_0 - \sum_{i=0}^{L} c\beta_i z_i, \qquad (5)$$

where β_i is the proportion of workers at level i, subject to the constraint that problem solvers at each level spend only one unit of time per capita answering problems, $[1 - F(\sum_{j=0}^{i-1} z_j)]h\beta_0 = \beta_i$.

In order to know how many layers of problem solvers the organization should have and how many workers there should be in each layer, a specific density of problems needs to be specified. Subsection C takes this step and analyzes the dependence of the organizational design

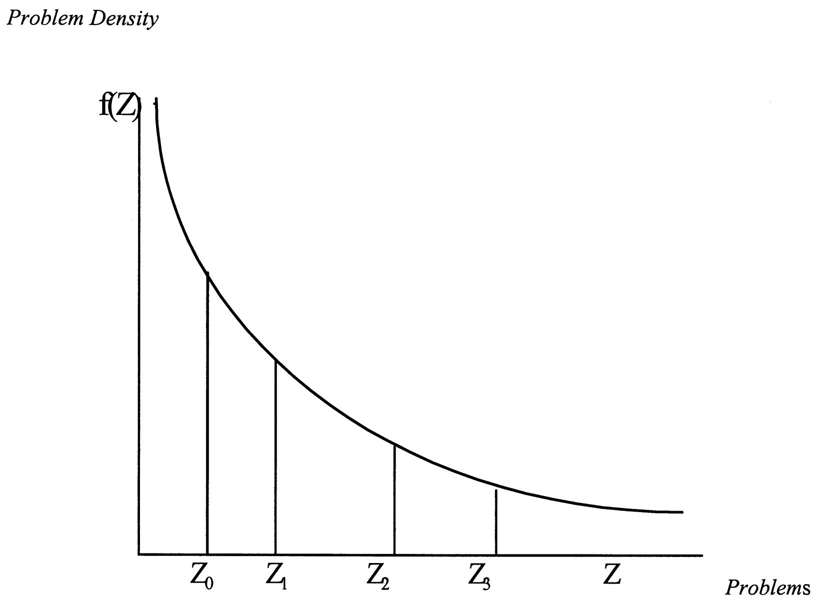

FIG. 2.—Organization of knowledge acquisition when communication is impossible. Production workers learn to solve problems $[0, Z_0)$ and present the rest successively to specialized problem solvers with knowledge $[Z_0, Z_1)$, $[Z_1, Z_2)$, etc.

characteristics (number of layers, workers per layer, and knowledge of workers in each layer) on the production and information technology. But first, I briefly reinterpret the model to apply it to the case in which problems differ in their difficulty rather than in their frequency.

C. *An Alternative Interpretation: Organization When Problems Differ in Their Difficulty*

Suppose that problems differ in their difficulty rather than in their frequency. The previous argument can be extended to show that the optimal solution has the same characteristics as the one above.

For notational convenience, assume that problems are uniformly distributed in the interval $\Omega = [0, 1]$. Production requires that problem $Z \in [0, 1]$ be drawn and solved. Let $c(Z)$ be the cost of learning a problem Z. Normalize this function so that the easiest problems (those with the lowest learning cost) correspond to the lowest values of Z, so that the function $c(Z)$ is nondecreasing. Then the cost of learning the set of problems in A is $\rho(A) = \int_A c(Z)dZ$, where the function $\rho(A)$ is a measure of the interval A.

Communication costs are

$$\beta_i t_i^h = \sum_{k\,:\,i\in l_k} \beta_k t_k^p \left[1 - \left(\bigcup_{l \prec_k i} A_l\right)\right] h;$$

net output is

$$y = \sum_{i=1}^{L} \left[\beta_i t_i^p \left(\bigcup_{k\in l_i} A_k\right) - \beta_i \rho(A_i)\right].$$

The same argument as in proposition 1 leads here to an optimal solution with specialization in problem solving or in production. The proof is analogous since the one before holds for a given allocation of knowledge A_i, independently of the fact that the cost of learning is now given by a more general measure. For a given allocation of knowledge, the organization problem is linear in the team sizes, and workers should be reallocated so that the class with the highest net output per worker has the largest number of workers and the rest are supporting it.

The argument in proposition 2 implies again here that knowledge should not overlap. Proposition 4 similarly goes through unchanged. The later production workers ask one class of workers, the fewer members this class needs to have in order to answer their questions, since there will be a smaller proportion of problems left to be solved.

Proposition 3 is the only one that needs to be reformulated. Since the size of a class of workers is smaller the later it is asked, output per worker is maximized if production workers learn the easiest problems and then successively ask those who know more and more difficult problems. The argument here proceeds as it did earlier: keeping team size constant, swap intervals of relatively harder problems learned by classes of workers at the start of the list for easier ones learned by other workers. Communication costs are now unchanged since all problems are equally frequent, but learning costs are minimized since the smaller classes (those higher up) are the ones learning the costliest problems. Thus this solution has a smaller number of workers learning the hardest problems. The principle guiding the organization now is increasing difficulty of the problems learned by workers further from the production floor.

Thus the organization can be characterized, analogously to the frequency-based organization, in the following way: (1) Workers specialize either in production or in problem solving. Only one class specializes in production. (2) Knowledge acquired by different classes does not overlap. (3) Production workers learn to solve the easiest problems and problem solvers learn to solve the harder ones. Moreover, the higher up in the list of production workers a problem solver is, the harder the problems she is able to solve. (4) The organization has a pyramidal structure, with each layer a smaller size than the previous one.

In what follows, we shall return to the frequency interpretation of the organization and study how changes in technology affect the organization of knowledge in production.

III. Technological Change and Organizational Design

A. *A Specific Model*

The model as presented may be characterized by three parameters: the cost of acquiring knowledge c, the cost of transmitting this knowledge h, and the predictability of the production process, understood as the extent to which "unexpected" problems are confronted by the organization.[5] In order to analyze the effects of technological changes on the organization, it is useful to make some assumption about the specific density of problems involved in production. I assume that the density of problems has the mathematically convenient exponential form with parameter λ. This parameter uniquely determines the characteristics of the production environment: a higher λ is always preferred since it implies a more "predictable" environment.

Because of the memoryless property of the exponential density, the number of layers (L) of problem solvers is unlimited in the absence of integer constraints. The value of the extra layer is given by the conditional probability that the problem solution is found in that layer given that it was not in the previous layers, and this is a constant, independent of how many layers the organization has.[6] In what follows, I obtain the solution when the organization is very large, so that the number of layers can be approximated as infinite.

The organizational problem is then

[5] More formally, process a is more predictable than process b if the cumulative distribution function of problems associated with b, $F_b(z)$, (first-order) stochastically dominates the distribution of problems associated with a, $F_a(z)$, i.e., $F_a(z) > F_b(z)$ for every z; in other words, the density function corresponding to b has a "thicker" tail.

[6] Add a one-worker layer (adding a layer is necessarily a discrete change) to an organization with total knowledge Z. Knowledge of this worker is $[Z, Z+a)$. She fully occupies her time (given constant returns to scale) and answers questions for $1/\{[1-F(Z)]h\}$ production workers, increasing their output by $F(Z+a)-F(Z)$. Thus the expected gross output gain is

$$\frac{F(Z+a)-F(Z)}{[1-F(Z)]h}=\frac{1-\exp(-\lambda a)}{h},$$

which is independent of Z. The marginal cost of a layer is the training cost plus the wage $ca+w$. The condition is independent of Z: if adding one layer is optimal, then it is always optimal to add an additional one.

$$\max_{z,\beta}\lim_{L\to\infty}\left[F\left(\sum_{i=0}^{L} z_i\right)\beta_0-\sum_{i=0}^{L} c\beta_i z_i\right] \tag{6}$$

subject to the constraint that $\sum_{i=0}^{\infty}\beta_i = 1$ and the time constraints on problem solvers, $1 - F(\sum_{j=0}^{i-1} z_j)h\beta_0 = \beta_i$, or, when $F(z)$ is exponential,

$$\exp\left(-\lambda\sum_{j=0}^{i-1} z_j\right)h\beta_0 = \beta_i. \tag{7}$$

We can use this set of constraints to eliminate the team size β_i from the optimization (6). Intuitively, the size of a layer is given by the proportion of workers asking questions to this layer (β_0) and by the knowledge of previous layers. From $\sum_i\beta_i = 1$, we can also eliminate β_0 from the optimization and write (6) as a function exclusively of the knowledge acquired by the workers at each layer. Then the problem of the organization is choosing the knowledge of workers at each level that maximizes output per capita:

$$y^* = \max_{z}\frac{F\left(\sum_{i=0}^{\infty} z_i\right) - cz_0 - \sum_{i=1}^{\infty} chz_i \exp\left(-\lambda\sum_{j=0}^{i-1} z_j\right)}{1 + h\sum_{i=0}^{\infty}\exp\left(-\lambda\sum_{j=0}^{i} z_j\right)}. \tag{8}$$

The first-order condition for z_0 is proportional to

$$f\left(\sum_{i=0}^{\infty} z_i\right) - c + \lambda ch\sum_{i=0}^{\infty} z_{i+1}\exp\left(-\lambda\sum_{j=0}^{i} z_j\right) + y^*h\lambda\sum_{i=0}^{\infty}\exp\left(-\lambda\sum_{j=0}^{i} z_j\right) = 0. \tag{9}$$

The marginal value of more knowledge acquired by production workers is the decrease in the learning costs of higher-learning workers since fewer of them are needed, and the increase in the production time permitted by a smaller amount of problem-solving time. The marginal cost is the marginal learning cost of these workers.

The first-order conditions for z_k for all $k > 0$ are proportional to

$$f\left(\sum_{i=0}^{\infty} z_i\right) - ch\exp\left(-\lambda\sum_{i=0}^{k-1} z_i\right) + \lambda ch\sum_{i=k}^{\infty} z_{i+1}\exp\left(-\lambda\sum_{j=0}^{i} z_j\right) + y^*h\lambda\sum_{i=k}^{\infty}\exp\left(-\lambda\sum_{j=0}^{i} z_j\right) = 0, \tag{10}$$

which has an analogous interpretation.

It is intuitive that, at the optimum, $z_k = z_{k+1}$ for $k > 0$, since the marginal cost and the marginal value of knowledge are independent of the level of workers.[7] This can in fact be verified,[8] and the solutions are

$$z_s^* = \frac{1}{\lambda}\ln\left(\frac{\lambda}{c} - \ln h\right), \tag{11}$$

$$z_w^* = \frac{1}{\lambda}\ln\left(\frac{h\lambda}{c} - h\ln h\right) = z_s^* + \frac{1}{\lambda}\ln h, \tag{12}$$

$$s = \frac{\lambda}{c} - \ln h, \tag{13}$$

and

$$y^* = 1 - \frac{c}{\lambda}\left[1 + \ln\left(\frac{h\lambda}{c} - h\ln h\right)\right], \tag{14}$$

where z_w^* is the length of the interval of problems that production workers can solve, whereas all the problem solvers learn to solve an interval of problems of equal length z_s^* (i.e., $z_i = z_s^*$ for all $i > 0$). Finally, $s = s_i = \beta_i/\beta_{i+1}$ is the "span of control" of each layer of problem solvers.

We can use expressions (11), (12), and (13) to obtain the ratio of production workers to problem solvers (r), the frequency of decision making by problem solvers (f), and the average "delay" (d) in obtaining the solution to a problem, understood as the expected number of layers of problem solvers involved in solving a given problem. The number of layers of the organization (L) is limited only by integer constraints. We shall approximate this limit by characterizing the number of layers involved in solving a given proportion of problems.

[7] It may seem that the marginal benefit of a layer must be smaller the higher the layer. In fact, the size of the layer is smaller, but the layer is always fully occupied in helping, so the marginal value of the layer, given by the conditional probability that the solution will be found there given that it was not found in previous layers, is constant.

[8] Conditions (9) and (10) imply that, when $z_i = z_j$ for all $i, j > 0$,

$$\exp(\lambda z_s) = \lambda z_s + 1 + \frac{\lambda}{c}y^*.$$

Substituting z_s and y^*, we have an identity. The other condition can be verified in a similar manner. An interior solution requires that $z_0 > 0$, which is true for parameter values such that $h[(\lambda/c) - \ln h] > 1$. This solution can be verified to be a maximum by substituting it in the second-order conditions.

B. Comparative Statics

We can now use these expressions to study the impact of changes in the three parameters of the model (h, c, and λ) on the design of the organization. The first two parameters, the cost of communication and the cost of acquiring knowledge, represent, as I shall argue, two different aspects of what is usually classified under the heading of "information technology." The third, λ, represents the complexity of the production process.

First, what is the impact of technologies, such as electronic networks and electronic mail, that reduce the cost of communication, allowing the knowledge of each problem solver to be more cheaply transmitted? The following proposition summarizes the impact of a reduction in h on organizational design.

Proposition 5. *Communication cost.*—A decrease in the cost of communication (h) has the following effects: (1) It increases the range of expertise of problem solvers (z_s) and reduces the range of expertise of production workers (z_w). As a consequence, the frequency of decision making by problem solvers, $1 - F(z_w)$, increases, and the frequency of decision making by production workers, $F(z_w)$, decreases. (2) It increases the span of control of problem solvers at each level (s). It has an ambiguous effect on the average delay required to find a solution (d) and on the average number of layers needed to solve a given proportion of problems ($\bar{L}$).

Proof. The result is immediate for z_s, z_w, and s from equations (11), (12), and (13). The frequency of decision making by production workers, $F(z_w)$, moves in the same direction as z_w since the density is unchanged. To study the effect on delay, obtain

$$\begin{aligned} Ed &= 1 \times P\{z_0 < Z < z_s + z_0\} + 2 \times P\{z_0 + z_s < Z < 2z_s + z_0\} + \cdots \\ &= \left[h\left(\frac{\lambda}{c} - \ln h - 1\right)\right]^{-1}, \end{aligned}$$

and $\partial Ed/\partial h$ cannot be unambiguously signed.

Second, L layers of problem solvers can solve a proportion of problems $1 - \delta$ given by

$$F(z_w + Lz_s) \equiv 1 - \frac{1}{h[(\lambda/c) - \ln h]^L} = 1 - \delta.$$

So that each δ determines a minimum number of layers $\bar{L}$ required to solve it,

$$\bar{L} = \left\lceil -\frac{\ln (\delta h)}{\ln [(\lambda/c) - \ln h]} \right\rceil, \tag{15}$$

where the *ceiling function* $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Although derivatives cannot properly be taken, since $\bar{L}$ is an integer, it is clear from taking derivatives of the argument of the ceiling function that a change in h could result in an increase or a decrease in the number of layers of problem solvers that are required to solve a proportion of problems $1 - \delta$. Q.E.D.

In the terms used by the popular press, the organization becomes flatter, and workers are less “empowered” as a result of an improvement in communication technology. The intuition for this result is as follows. First, as communication becomes cheaper, relying on problem solvers is cheaper, and, as a consequence, it is optimal for each production worker to acquire less knowledge. Moreover, each problem solver can communicate solutions to a larger team, so that the span of control of problem solvers increases. The ambiguous effect on delay and the number of layers needed to solve a given proportion of problems is due to the opposite changes in the knowledge of production workers and problem solvers (z_w, z_s): it is more likely that workers need to ask (increasing delay and the number of layers involved in solving a given proportion of problems); but since problem solvers each know more, conditional on being asked, they are likely to obtain the answer sooner.

Second, the cost of acquiring knowledge (c), as understood here, is affected by changes such as the introduction of expert systems and electronic diagnostics: each worker can solve, for a given investment in acquiring knowledge, a larger proportion of problems. For example, a machine operator can solve more problems for a given investment in learning if the machine is fitted with a diagnostic system. The following proposition summarizes the effects of a reduction in c in this context.

Proposition 6. *Cost of acquiring knowledge.*—A reduction in the cost of acquiring knowledge (c) has the following effects: (1) It increases the range of expertise of both problem solvers (z_s) and production workers (z_w). As a consequence, the frequency of decision making by production workers, $F(z_w)$, increases and the frequency of decision making by problem solvers, $1 - F(z_w)$, decreases. (2) It increases the span of control of problem solvers (s). It also reduces the average delay necessary for finding a solution (d) and does not increase, and may reduce, the number of layers necessary to solve a given proportion of problems $\bar{L}$.

Proof. As before, the result is immediate for z_s, z_w, and s from equations (11), (12), and (13). Again, the frequency of decision making by production workers, $F(z_w)$, moves in the same direction as z_w since the density is unchanged. Problem solvers solve the problems not solved by production workers, $1 - F(z_w)$. That the number of layers required to solve a given proportion of problems ($\overline{L}$) is nondecreasing in c can be seen immediately from equation (15) in the previous proof, defining this quantity. The delay $\{h[(\lambda/c) - \ln h - 1]\}^{-1}$ is reduced as more solutions are encountered close to the production floor, and each problem solver, if asked, has a higher probability of knowing the answer. Q.E.D.

The organization has become "flatter" but workers are now more, rather than less, "empowered." Production workers can acquire knowledge more cheaply, so they ask relatively fewer questions. This increases the span of control of each problem solver, reduces the number of layers of problem solvers required to solve a given proportion of problems, reduces the delay needed to obtain solutions to problems, and decreases the frequency with which problem solvers intervene in the production process.

Finally, changes in the density function as indexed by λ are (the inverse of) changes in the "complexity" of the production process. A more complex production process is one in which problems farther out in the tails are more likely to be confronted. The following proposition summarizes the effects of a change in λ.

Proposition 7. *Predictability.*—A decrease in the predictability (increase in the complexity) of the production process (λ) has the following effects: (1) It may increase or decrease the range of expertise of problem solvers (z_s) and production workers (z_w). However, it unambiguously increases the frequency of decision making by problem solvers, $1 - F(z_w)$, and decreases the frequency of decision making by production workers, $F(z_w)$. (2) It reduces the span of control of problem solvers (s). It increases the average delay necessary to find a solution (d) and does not reduce, but may increase, the number of layers of the organization required to solve a given proportion of problems ($\overline{L}$).

Proof. As before, the results for z_s, z_w, and s are immediate. The nonincreasing impact of a change in λ on $\overline{L}$ follows as before from its definition in equation (15). To see that the effect on the frequency of decision making is unambiguous, note that

$$1 - F(z_w) = \left[h\left(\frac{\lambda}{c} - \ln h\right)\right]^{-1}.$$

The effect on the average delay follows unambiguously from

$$Ed = \left[h\left(\frac{\lambda}{c} - \ln h - 1\right)\right]^{-1}.$$

Q.E.D.

A change in the density function may be an increase or a decrease in the marginal value of knowledge since the density rotates around as

λ shifts. This is why the effect on the range of expertise is ambiguous. But the effect on the frequency of decision making is clear: as the production process becomes more complex, production workers need to rely more often on problem solvers. As a consequence, their span of control decreases, and the organization has more layers of a minimal size or greater.

Results on the relation between the predictability of the production process and aspects of organizational design were previously available only in Athey et al. (1994). They assumed that the intrinsic value of a decision by different employees in different states of the world is different and that the states in which managers have an intrinsic advantage are relatively infrequent. Moreover, the organization is restricted to one manager and one worker, which implies that the model has no implications for the allocation of workers to layers and the number of layers in the organization.

Their results are not incompatible with the ones found here, but they are substantially different. More complexity, in their case, means that more weight is put on states in which the worker is intrinsically a worse performer, which implies that he can take on more tasks and the manager fewer. Thus more complexity implies under most circumstances more discretion for production workers and less discretion for supervisors (this is ambiguous in the analysis I present), whereas the change in the frequency of decision making could go in any direction (here the frequency of decision making by problem solvers unambiguously increases). Their result keeps the number of layers and the spans of control constant.

The analysis presented here does not impose a priori any difference in the intrinsic value of the performance of managers and workers in a given state. A more complex production process is simply one that puts more weight on unusual states. As production becomes more complex, the analysis allows the number of layers involved in solving a certain proportion of problems to increase and the span of control to decrease at the same time the task assignment changes. This increases the value of learning by production workers (as Athey et al. argue) but decreases the value of the marginal knowledge since the marginal probability of getting solutions with more knowledge may be lower. Finally, more complexity implies that production workers make decisions less often (and problem solvers more) since the increase in their learning, if it exists, is never sufficient to compensate for the larger weight of unusual problems.

The three propositions in this section have analyzed the impact of technological changes on organizational design. Figure 3 presents these results. It shows the proportion of workers assigned to each level (β) and the probability that a problem is solved up to that level

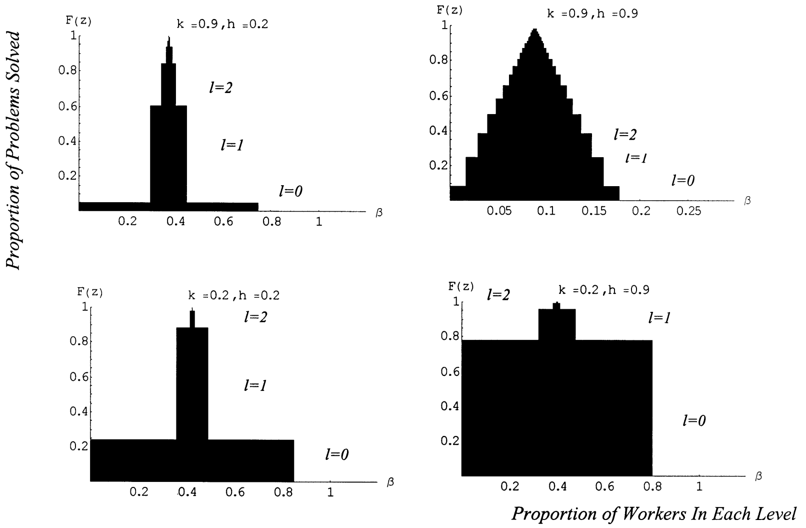

FIG. 3.—Optimal organizational forms when knowledge is nonoverlapping as a function of communication cost (h) and net learning cost ($k = c/\lambda$). Organizations are characterized by the problem-solving ability of workers up to a level ($F(Z)$) and proportion of workers (β) assigned to each level (l). The number of layers when knowledge is nonoverlapping is unlimited, but only a limited number of layers can be shown graphically since they become increasingly small. Note that in the graph in the upper left corner, production workers acquire no knowledge (they ask every problem they confront), even though the graph shows positive but small knowledge in order to make these workers visible.

($F(\sum_{j=0}^{i} z_j)$) as a function of communication cost (h) and the ratio of learning cost to hazard rate, $k = c/\lambda$, or "net" cost of acquiring knowledge. As one moves from left to right, communication costs increase. As one moves upward, net learning cost increases. Improvements in any of the "information technology" parameters increase the average span of control of problem solvers (the ratio of the length of one level to the previous one). However, they have different effects on the range of expertise or "discretionality" of production workers and employees: while reductions in the cost of communication (h) (as one moves left) decrease the proportion of problems solved by production workers, as they rely more on problem solvers, reductions in the net cost of acquiring knowledge (c/λ) (as one moves down) increase the proportion of problems solved by the production workers.

IV. Extension: Optimal Organization When Knowledge of Different Layers Must Overlap

The organization proposed in the previous sections does not require any overlap in knowledge acquired by workers at different levels. This is true in many real-world examples. The knowledge of a production engineer often does not encompass the more detailed and practical knowledge of a machine operator. In other cases, however, the knowledge of the more knowledgeable worker encompasses the knowledge of the least knowledgeable one. This is particularly true when knowledge is tacit, so that most knowledge must be acquired on the job, in the form of learning by doing.

This section analyzes the robustness of the results presented before to the extra constraint that knowledge be overlapping. It shows that the main characteristics of the solution, such as the pyramidal shape of the optimal organization and the existence of specialization between production and transmission of knowledge, remain unchanged. The most notable difference is that now there are strongly diminishing returns to the number of layers (since adding an extra layer implies adding workers who acquire knowledge that they in part never use), and thus the number of layers at the optimum is always relatively small.

Consider, as in Section II, an organization with L classes of workers, whose time is employed in production and in communication of problem solutions, but whose knowledge must overlap, so that, for all i, $A_i \subset A_{i+1}$. Given this restriction, the time constraint of problem solvers can be written identically as before:

$$\beta_i t_i^h = \sum_{k:\, i \in l_k} \beta_k t_k^p \left[1 - F\left(\bigcup_{l \prec_k i} A_l\right)\right] h, \quad \text{for } i = 1, \ldots, L. \tag{16}$$

The output of workers of class i is still

$$y = \sum_{i=1}^{L} \left[\beta_i t_i^p F\left(\bigcup_{k \in l_i} A_k\right) - c\beta_i \mu(A_i)\right]. \tag{17}$$

Formally, the problem is identical to the one formulated previously, subject to the restriction that $A_i \subset A_{i+1}$. All the results obtained previously, concerning specialization, information flow, management by exception, and pyramidal shape, go through. The only result that does not hold, since we have restricted it in that direction, is the one concerning overlapping knowledge (proposition 2); obviously, the knowledge of each layer encompasses, by assumption, the knowledge of the previous layer.

Proposition 8. *Organization with overlapping knowledge.*—When, because of the need for learning by doing, knowledge of one layer must encompass knowledge of the previous ones, the optimal organization has the following characteristics: (1) Workers specialize either in production or in problem solving. (2) Production workers learn to solve the most common problems and ask specialized problem solvers to solve the rest. Moreover, information always flows in the same direction, and the flow is "vertical": problem solvers are asked according to their position in the problem density, so that those asked first know about problems associated with the highest density. (3) The organization has a pyramidal structure, with (possibly) several successive layers of problem solvers of a decreasing size. Those in the highest layer acquire the most knowledge, and their knowledge encompasses the most unusual problems.

Proof. See the proofs of propositions 1–4. Q.E.D.

We can now go on to consider, as in Section III, a specific density and study the comparative statics changes in organizational design derived from changes in the information technology and production parameters. As previously, let $\beta_0, \beta_1, \dots, \beta_L$ be the share of workers at each level, and let z_i be the length of the knowledge interval of worker i (note that $z_i > z_{i-1}$ for all i in this setting by assumption). Expected net output of a firm with L layers of problem solvers is

$$y = F(z_L)\beta_0 - \sum_{i=0}^{L} c\beta_i z_i, \tag{18}$$

subject to the constraints that all problem solvers spend one unit of time helping production workers, $[1 - F(z_{i-1})]h\beta_0 = \beta_i$; that learning be always nonnegative, $z_i \geq 0$; and that all the workers be assigned to some level $\sum \beta_i = 1$.

We can substitute in the constraints on the number of workers in each level $\exp(-\lambda z_{i-1}) = \beta_i/h\beta_0$ to write the problem of the firm as choosing the number of layers, the number of workers in each layer, and the amount of knowledge they acquire to maximize net output per capita:[9]

$$\max_{L,z,\beta} F(z_L)\beta_0 + k\sum_{i=1}^{L} \beta_{i-1}\ln\beta_i - k\sum_{i=1}^{L}\beta_{i-1}\ln h\beta_0 - c\beta_L z_L \tag{19}$$

subject to nonnegativity constraints and the constraint $\sum_{i=0}^{L}\beta_i = 1$.

Solving the firm's optimum requires maximizing the firm's production over the number of levels. Since output is a discrete function of L, I first obtain the optimal output for each L as a function of c, λ, and h and then maximize numerically over L. Since all the parameters can

[9] The term k is defined to be $k = c/\lambda$.

be reduced to simple functions of two variables ($k = c/\lambda,\ h$), it is possible to solve the problem numerically and represent the solution graphically.[10]

Taking the first-order conditions of this maximization and using the definition $s_i = \beta_i/\beta_{i+1}$, we obtain two sets of solutions, depending on whether production workers acquire some knowledge. After some manipulation of the first-order conditions, the interior solution for a given number of levels of problem solvers L is described by the second-order recursion[11]

$$(1): \quad s_0 + \frac{1}{\beta_0} - \frac{1}{k} = \ln s_1, \tag{20}$$

$$(i): \quad s_i - s_{i-1} = \ln s_{i+1}, \quad \forall\ i = 1,\ \dots,\ L-1, \tag{21}$$

$$(L): \quad s_{L-1} - s_{L-2} = -\ln hk, \tag{22}$$

and

$$(l+1): \quad 1 + \frac{1}{s_0} + \frac{1}{s_0 s_1} + \cdots + \frac{1}{s_0 \cdots s_{L-1}} = \frac{1}{\beta_0}. \tag{23}$$

Thus the essential features of the organization presented in Section III persist when knowledge is overlapping. Figure 4 shows the same two variables presented in figure 3 for the purpose of comparing the two cases. The figures for small h show a substantially similar configuration in both cases. The only stark difference between figures 3 and 4 exists for high h and k, in the upper right corner. Note that the spans of control (s) are as low as in the nonoverlapping solution, and the amount of knowledge acquired by those layers that exist is roughly as low. However, only two layers exist. A solution like the one in figure 3 (a large number of levels) is impractical since adding a layer of workers who do not use most of the knowledge they learn is adding a particularly large deadweight loss when knowledge is expensive.

[10] A previous version, available from the author, obtained the full numerical solution to this problem.

[11] The corner solution that has $z_0 = 0$ (workers learn nothing) is described by the same equations except that eq. (20) becomes $s_1 = 1/h$ and eq. (21) for $i = 2$ becomes

$$(2): \quad \frac{1}{\beta_0} - (1+h)s_2 - \frac{1}{k} - 1 = \ln s_2 + (h+1)\ln s_3.$$

The rest of the second-order recursion and the constraint on the total number of workers (eqq. [21], [22], and [23]) remain the same as in the interior solution.

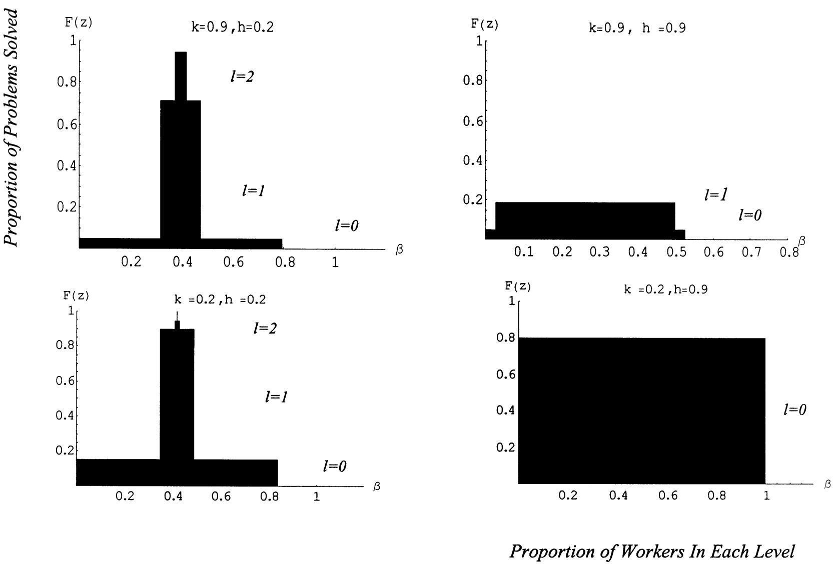

FIG. 4.—Optimal organizational forms when knowledge is overlapping as a function of communication cost (h) and net learning cost ($k = c/\lambda$). Organizations are characterized by the problem-solving ability of workers up to a level ($F(Z)$) and proportion of workers (β) assigned to each level (l). Note that in the two graphs on top, production workers acquire no knowledge even though the graphs show positive but small knowledge in order to make these workers visible.

V. Implications and Discussion

This paper has examined an organization whose aim is to structure the acquisition of knowledge so as to economize on learning and communication costs. This section explores the implications of the theory and to what extent they correspond to interesting economic phenomena.

A. Knowledge-Based Hierarchies and "Management by Exception"

We have found that, when matching problems with experts is very costly, the optimal organization of productive knowledge has the features of a hierarchy. Production workers acquire knowledge about the most common problems they confront, whereas problem solvers specialize in acquiring and transmitting knowledge in the form of directions about what to do when the worker confronts a problem she does not know. The optimal organization has a pyramidal shape, with the production workers at the base, and fewer workers acquire knowledge about exceptional problems. Communication flows vertically, from those who know the more common problems to those who know increasingly exceptional ones. Depending on the technology available for acquiring knowledge, knowledge of higher-level workers may or may not encompass the knowledge of lower-level employees.

Examples of this kind of division of knowledge are widespread. Alfred Sloan (1924, p. 195), in describing his work, claimed that "we do not do much routine work with details. They never get up to us. I work fairly hard, but it is on exceptions …, not on routine or petty details." On the shop floor of a production plant, machinists deal with most of the problems involved in the operation of the machines. A stoppage that presents unusual problems may require the attention of a mechanical supervisor, and only in truly unusual circumstances is the presence of a production engineer required. In a cardiac care room, interns and residents are in direct, continuous contact with patients. Physicians intervene only when residents encounter a problem that is sufficiently unusual.

B. Information Technology and Organizational Design

Two of the parameters that play a role in the theory suggested here, the cost of communicating knowledge among workers and the cost of acquiring knowledge, are altered by recent improvements in information technology. First, expert systems and the codification allowed by computers reduce the cost of acquiring the knowledge necessary to solve a given proportion of possible problems. The model predicts that such an expert system would increase the scope of decision making by lower-level workers, increase the span of control of supervisors, increase the ratio of production workers to problem solvers, and reduce the number of layers of workers with specialized knowledge required. Second, current innovations in information technology also translate into smaller costs of communicating knowledge among workers. The theory presented here predicts that reductions in the cost of communicating knowledge also increase the ratio of production workers to problem solvers, but they increase the reliance of these workers on problem solvers, decreasing their own scope of decision making.

The evidence on the changes that organizations are experiencing in these three dimensions as a result of the decrease in the costs of information technology is spotty. Some micro research points to an increase in the span of control of managers (e.g., Batt 1996). Osterman (1996) presents evidence showing a reduction in the number of layers in hierarchies in diverse sectors, such as insurance, telecommunications, and automobiles. There is also a large amount of discussion suggesting an increase in the discretionality ("empowerment") of lower-level workers. Brickley, Smith, and Zimmerman (1996, p. 232) agree that recently "there has been a shift towards granting employees broader decision authority and less specialized task assignment." The evidence seems consistent with the interpretation that information technology has allowed workers cheaper access to knowledge (a decrease in c): firms assign broader decision-making ability to lower-level employees as a consequence of their access to cheaper knowledge. As a consequence, they need their supervisors less often, increasing their span of control and decreasing the length of the hierarchy.

C. Vertical Integration: Outsourcing Problem Solutions When Problems Are Too Infrequent

The theory presented here has ignored integer constraints. This allows the organization to have layers of infinitely small size, specializing in solving problems with a remote probability of occurring. Relaxing this assumption makes immediately apparent a rationale for vertical integration and outsourcing: it is not worth having in-house the ability to solve extremely exceptional problems. The role of outside consultants in such a world would be to solve those problems that happen so infrequently that it is not optimal to have inside staff solving them.

Introducing in the model in Section III the restriction that, to remain in the organization, layers must have a given minimum number of workers (e.g., one) leads immediately to implications on vertical scope. Suppose that the number of production workers P in the organization is exogenously given, so that the total number of workers involved in the production process is P/β_0. The number of workers in layer n is $\beta_n P/\beta_0 = P/s^n$, where the span of control s is given by the expression in (13). The number of workers in layer n is larger than one if $P/s^n > 1$ or, equivalently, if $\ln P/\ln s > n$. Call $\bar{n}$ the last layer n for which this condition holds; abusing the notation (since n is discrete), we can approximate this number as $\bar{n} \approx \ln P/\ln s$. The proportion of problems that are too uncommon to merit one full member of the organization to deal with them is $O = 1 - F(z_w + \bar{n}z_s)$. Substituting in the values of z_w, z_s, and $\bar{n}$, we have $O \approx 1/shP$.

It is now easy to analyze the effects on the proportion of problems outsourced of changes in h, c, and λ in an organization with a given number of production workers. First, $dO/dh \approx (1 - s)/h^2Ps^2 < 0$ since $s > 1$. This result means that the proportion of problems outsourced increases when the cost of communication decreases. Intuitively, a decrease in the cost of communication (h) implies that more problems are solved by higher layers, which at the same time have become too small to be kept in-house.

The effect of a change in the cost of acquiring knowledge (c) is the opposite since $dO/dc \approx \lambda/c^2hPs^2 > 0$. This means that fewer problems are outsourced when acquiring knowledge is cheaper. Intuitively, this follows from the fact that a decrease in c leads to the acquisition of more knowledge by the lower layers of the organization. Finally, an increase in predictability of the production process (λ) can be shown similarly to lead to a decrease in the proportion of problems outsourced. Thus we expect organizations engaged in simpler activities to solve a larger proportion of their problems in-house.

D. Predictability of Production and Organizational Design

The theory predicts that more complex production processes increase the need to rely on higher-level problem solvers, reducing their span of control.

Some evidence supporting these implications has been suggested by Jay Galbraith and other researchers in the field of contingency theory.[12] In particular, organization theorists have found a relation between the ratio of supervisors to workers and both the predictability of the production process and the skill of production workers (Galbraith 1977, pp. 35, 36). In the framework I propose, as a process becomes more predictable or as workers closer to the production floor become more skilled, lower-level workers are able to solve a larger proportion of problems and need to use the help of specialized problem solvers less often.

E. Coordination Costs Limit Specialization

The theory presented here is directly interpreted as a motivation for a hierarchical division of knowledge. Alternatively, it could be interpreted as a metaphor for the problem of specialization. The model builds on two points already present in some of the literature on specialization. First, as Rosen (1983) points out, a motor for specialization is the fact that learning involves a fixed cost, independent of its utilization. As a consequence, its economic return increases with the intensity of its use. The second building block answers the question, What limits the amount of specialization that takes place in equilibrium? Becker and Murphy (1993) point out the essential trade-off between coordination costs and specialization. They provide some evidence of the importance of this trade-off in avoiding the monopolistic consequence that the statement that "the division of labor is limited by the extent of the market" would imply.

[12] This theory proposes that the structure of an organization must respond to the external environment. Galbraith suggests that managers' role is to handle exceptions as a consequence of the need to make decisions that cannot be made according to the usual rules (Galbraith 1973, pp. 10–11).

The model I present builds on these insights. By referring explicitly to a certain kind of knowledge (tacit problem-solving knowledge) and the content of communication, I am able to obtain inferences on the extent of specialization, but also on the frequency of communication and on the proportion of workers specializing in each task.

F. The Market for Knowledge

The obstacle that the efficient functioning of the market for knowledge is usually thought to confront is that acquiring knowledge involves only a fixed cost. This introduces increasing returns to scale in production, with the consequence that competition among those who have knowledge will drive the price of knowledge to zero. The snag in this argument, as this model makes clear, is that communicating knowledge is not free. If communication is costly, then those competing in this market face a time constraint that ensures that the price of knowledge transfer is equal to the average cost of learning and communicating solutions.

In the context of this model, the market for knowledge may be thought of as formed by profit-maximizing production workers who acquire some knowledge and demand answers to the problems they cannot solve, and problem-solving workers who, in exchange for a fee per question asked, acquire specialized knowledge and provide answers to the problems that production workers cannot solve. A more knowledgeable problem solver can charge a higher fee per question since workers are willing to pay a higher price per question in exchange for increasing the probability that the worker asked solves the problem so that they can avoid going on to other problem solvers. The market equilibrium must balance the demand for and supply of knowledge and the demand for and supply of questions and answers, and make all workers indifferent about being production workers or problem solvers at any level. A previous version of this paper (available from the author) shows that the decentralized equilibrium exists and solves for the equilibrium fees and wages per question.

VI. Concluding Remarks

This paper has developed a formal model of the role of hierarchical organization in solving problems encountered in production. In the spirit of Alchian and Demsetz (1972), a hierarchy is not defined by “some superior authoritarian directive or disciplinary power.” Instead, the role of supervisors is to transmit their knowledge about exceptional problems to production workers in the form of directions.

The analysis has made two simplifying assumptions: that workers are homogeneous and that problem flow is observable. If workers have different learning and communication abilities, designing the optimal organization involves assigning workers to positions in the hierarchy and obtaining equilibrium skill-wage functions, as similarly discussed in Rosen (1982). Second, if problems are unobservable to firms, firms must design incentive systems to ensure that workers deal with the the right problems rather than over- or underreferring them to other layers. These two problems need to be addressed by future work (Garicano and Santos 2000).

Appendix

Proof of Proposition 1

Let $\beta_i t_i^p = T_i^p$. Then output can be written as

$$y = \sum_{i=1}^{L}\left[T_i^p F\left(\bigcup_{k\in L_i} A_k\right) - c\beta_i\mu(A_i)\right], \tag{A1}$$

with helping costs

$$T_i^h = \sum_{k:\, i\in s_k} T_k^p\left[1 - F\left(\bigcup_{l\prec_k i} A_l\right)\right] h. \tag{A2}$$

Since it must be optimal to spend all time (it is always productive), $t_i^h + t_i^p = 1$. Substitute $T_i^h = \beta_i(1 - t_i^p) \equiv \beta_i - T_i^p$. Then the system of constraints (A2) is

$$T_i^p + \sum_{k:\, i\in s_k} T_k^p\left[1 - F\left(\bigcup_{l\prec_k i} A_l\right)\right] h = \beta_i \quad \text{for } i = 1, \ldots, L. \tag{A3}$$

For any given assignment of knowledge to classes A_i and of workers to classes β_i, this is a system of L linear equations with L unknowns T_i^p for $i = 1, \ldots, L$. This implies first that an allocation of knowledge and workers to classes *uniquely* determines the time that each class spends in production and the time it spends solving problems for other classes. Second, the solution of this system of equations determines T_i^p as a *linear* combination of the β_i's:

$$T_i^p = \rho_{i1}\beta_1 + \cdots + \rho_{iL}\beta_L = \boldsymbol{\rho}_i'\beta.$$

Substituting this expression for T_i^p in (A1) and calling $\boldsymbol{\rho}_i$ the vector of coefficients just obtained $(\rho_{i0}, \rho_{i1}, \ldots, \rho_{iL})$, we obtain the following program:

$$y = \max_{\beta}\sum_{i=1}^{L}\left[F\left(\bigcup_{k\in L_i} A_k\right)\boldsymbol{\rho}_i'\beta - c\beta_i\mu(A_i)\right], \tag{A4}$$

where $F(\bigcup_{k\in L_i} A_k)$ is a parameter for a given allocation of knowledge. This maximization can be interpreted in the following way. Each of the $T_i^p = \boldsymbol{\rho}_i'\beta$ is one of L productive processes, which are feasible linear combinations of the β's given by the vectors of coefficients $\boldsymbol{\rho}_i$ and the terms in A. Substitutability within each productive process is linear.

Then, optimizing is equivalent to choosing the best productive process and rearranging the β's so as to make the corresponding time in that process T_i^p as large as possible. The constraints are that $\sum_{i=1}^{L}\beta_i = 1$ and that T_i^p be between zero and β_i, that is, $\beta_i \geq \boldsymbol{\rho}_i'\beta \geq 0$. The result of this linear optimization has $T_i^p = \beta_i$ for one i and $T_j^p = 0$ and $T_j^h = \beta_i$ for all j different from i. These L equations reduce to $L - 1$ independent equations of the form $\boldsymbol{\rho}_j'\beta = 0$ since the equation $\beta_i = \boldsymbol{\rho}_i'\beta$, corresponding to $T_i^p = \beta_i$, is necessarily implied by the others from the system (A3). Intuitively, workers of team i cannot spend any time helping other workers, given that the other workers are not engaging in production at all. The $L - 1$ equations $\boldsymbol{\rho}_j'\beta = 0$ plus the equation $\sum_{i=1}^{L}\beta_i = 1$ deliver unique values of β_i for which only one layer specializes in production and all the rest are supporting it. Q.E.D.

References

Alchian, Armen A., and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *A.E.R.* 62 (December 1972): 777–95.

Athey, Susan; Gans, Joshua; Schaefer, Scott; and Stern, Scott. "The Allocation of Decisions in Organizations." Research Paper no. 1322. Stanford, Calif.: Stanford Univ., Grad. School Bus., 1994.

Batt, Rosemary. "From Bureaucracy to Enterprise? The Changing Jobs and Careers of Managers in Telecommunications Service." In *Broken Ladders: Managerial Careers in the New Economy,* edited by Paul Osterman. New York: Oxford Univ. Press, 1996.

Becker, Gary S., and Murphy, Kevin M. "The Division of Labor, Coordination Costs, and Knowledge." In *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education,* 3d ed., by Gary S. Becker. Chicago: Univ. Chicago Press, 1993.

Beggs, Alan W. "Queues and Hierarchies." *Rev. Econ. Studies,* in press.

Bolton, Patrick, and Dewatripont, Mathias. "The Firm as a Communication Network." *Q.J.E.* 109 (November 1994): 809–39.

Brickley, James A.; Smith, Clifford W., Jr.; and Zimmerman, Jerold L. *Organizational Architecture: A Managerial Economic Approach.* Chicago: Irwin, 1996.

Calvo, Guillermo A., and Wellisz, Stanislaw. "Supervision, Loss of Control, and the Optimum Size of the Firm." *J.P.E.* 86 (October 1978): 943–52.

Crémer, Jacques. "A Partial Theory of the Optimal Organization of a Bureaucracy." *Bell J. Econ.* 11 (Autumn 1980): 683–93.

Galbraith, Jay R. *Designing Complex Organizations.* Reading, Mass.: Addison-Wesley, 1973.

———. *Organizational Design.* Reading, Mass.: Addison-Wesley, 1977.

Garicano, Luis, and Santos, Jesús. "Referrals." Working paper. Chicago: Univ. Chicago, Grad. School Bus., July 2000.

Geanakoplos, John, and Milgrom, Paul. "A Theory of Hierarchies Based on Limited Managerial Attention." *J. Japanese and Internat. Econ.* 5 (September 1991): 205–25.

Hayek, Friedrich A. von. "The Use of Knowledge in Society." *A.E.R.* 35 (September 1945): 519–30.

Krasa, Stefan, and Villamil, Anne P. "Optimal Multilateral Contracts." *Econ. Theory* 4 (March 1994): 167–87.

Osterman, Paul, ed. *Broken Ladders: Managerial Careers in the New Economy.* New York: Oxford Univ. Press, 1996.

Qian, Yingyi. "Incentives and Loss of Control in an Optimal Hierarchy." *Rev. Econ. Studies* 61 (July 1994): 527–44.

Radner, Roy. "Hierarchy: The Economics of Management." *J. Econ. Literature* 30 (September 1992): 1382–1415.

———. "The Organization of Decentralized Information Processing." *Econometrica* 61 (September 1993): 1109–46.

Radner, Roy, and Van Zandt, Timothy. "Information Processing in Firms and Returns to Scale." *Annales d'Economie et de Statistique,* nos. 25–26 (January–June 1992), pp. 265–98.

Rosen, Sherwin. "Authority, Control, and the Distribution of Earnings." *Bell J. Econ.* 13 (Autumn 1982): 311–23.

———. "Specialization and Human Capital." *J. Labor Econ.* 1 (January 1983): 43–49.

Sloan, Alfred. "The Most Important Thing I Ever Learned about Management." *System* 46 (August 1924).

Van Zandt, Timothy. "Organizations with an Endogenous Number of Information Processing Agents." In *Organizations with Incomplete Information: Essays in Economic Analysis,* edited by Mukul Majumdar. Cambridge: Cambridge Univ. Press, 1998.

———. "Real-Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents." *Rev. Econ. Studies* 66 (July 1999): 633–58.