RAND Journal of Economics
Vol. 40, No. 1, Spring 2009
pp. 190–208

Coordination and delay in hierarchies

Andrea Patacconi*

This article studies hierarchical organizations where concerns for fast execution are important and employees must be coordinated to avoid wasteful duplications of effort. Simple conditions are provided for the time spent on coordinating subordinates to be increasing and the span of control to be decreasing as one goes up the hierarchy, with equalities holding if delay is all that matters. When returns to specialization are substantial, the span of control also tends to widen and the hierarchy to flatten as urgency increases. The model suggests that concerns for fast execution may be key in explaining recent trends toward decentralization and delayering in firms.

1. Introduction

■ One of the fundamental tasks facing management is to ensure that the activities within an organization are properly coordinated. This article contributes to the understanding of this issue by studying how concerns for fast execution affect organizational choices, in particular the allocation of coordination responsibilities across hierarchical levels.

Considerations of delay have long been recognized as key in shaping organizations. Rajan and Wulf (2006) report that Jeffrey Immelt, the CEO of General Electric, motivated the decision to shorten the chain of command at GE as follows: "The reason for doing this is simple—I want more contact with the financial services teams. . . . With this simplified structure, the leaders of these four businesses will interact directly with me, enabling faster decision making and execution."[1] Boeing is another case in point. According to Galbraith (1977), "after 1964 the problem facing Boeing was not to establish a market but to meet the opportunities remaining as quickly as possible. . . . Now a delay of a few months would result in canceled orders and fewer sales." Galbraith reports that, to respond to competitive time pressure from Douglas, Lockheed, and the British-French Concorde, Boeing was forced to drastically reduce the time devoted to product development and design. The initial product development phase, for instance, was compressed from four years on the 727 to four months on the 747. As one would expect, this

* University of Oxford; andrea.patacconi@economics.ox.ac.uk.

This article is based on Chapter 1 of my D.Phil. dissertation submitted to Oxford University. I am particularly indebted to Meg Meyer and John Quah for their support and advice. I also thank the editor, Mark Armstrong, and Karen Croxson, Wouter Dessein, Mathias Dewatripont, Erik Eyster, Ian Jewitt, Clare Leaver, Niko Matouschek, David Myatt, John Roberts, Armin Schmutzler, Dimitri Vayanos, and seminar participants at the Winter Meeting of the Econometric Society (Istanbul, 2005), ESSET (Gerzensee, 2006), Royal Holloway, Warwick, Queen Mary, and Duke for helpful remarks. Two anonymous referees provided extremely detailed and constructive suggestions that greatly improved the article. Financial support from the British Academy is gratefully acknowledged. All remaining errors are mine.

[1] Rajan and Wulf quote this passage from a General Electric press release titled "GE Announces Reorganization of Financial Services; GE Capital to Become Four Separate Businesses," July 26, 2002.

exacerbated coordination problems and created information overload. Boeing's primary response to these problems was to allocate "more resources, liaison and task forces, to coordination through local relations. . . . The lateral relations allowed a temporary decentralization of influence. The organization was still a functionally dominant one with an integrating role between functions" (Galbraith, 1977).

We study how concerns for fast execution interact with organizational choices in a simple model of an organization that must carry out a task. The task could be to build an airplane, to develop software, or to test a large number of chemical compounds in search of new drugs. For expositional ease, however, we will often refer to the task simply as information to be processed. We distinguish between managers, whose job is to plan and coordinate the allocation of duties among their direct subordinates, and workers, who engage in production at the bottom of the hierarchy. Coordination is important to reduce "the almost inevitable duplication in information processing that accompanies the use of multiple managers" (Geanakoplos and Milgrom, 1991). Specifically, we posit that by spending time on coordination activities such as work scheduling and the exchanging of information with colleagues, managers can reduce wasteful overlapping between tasks and thus the amount of information that must be processed by the organization. Examples include the chemical department at Du Pont, which by the end of 1920 took on the responsibility of coordinating the different manufacturing departments "so that overlapping of the research programs . . . may be avoided as completely as possible" (Chandler, 1990), and the coordination department at Jersey, which was responsible for advising the board so as "to avoid any needless duplication of equipment" (Chandler, 1962).[2]

Multilayer hierarchies can arise in this model for two reasons. First, when a task is delegated to multiple subordinates, a finer division of labor results, which may induce workers to specialize and thereby become more efficient. Second, delay may be reduced through parallel information processing. The main restriction placed on the hierarchy is that it must be balanced, so that all the managers *at the same level* must have the same number of subordinates (span of control) and spend the same time coordinating. This is essentially equivalent to assuming that tasks are divided evenly among subordinates, so that all the managers at the same level face exactly the same problem.

Production and coordination activities are assumed to be costly both in terms of wages and delay. The objective of the organization is to minimize a weighted average of the wage bill and the cost of delay. Thus, the weight attached to delay captures in a simple way the need for fast execution, or "urgency." The choice variables are the number of layers L in the organization, the time $\mathbf{t} = (t_1, \ldots, t_{L-1})$ managers spend coordinating at each layer of the hierarchy, and the span of control $\mathbf{s} = (s_1, \ldots, s_{L-1})$ at each layer of the hierarchy. We stress that, as in the information-processing literature, no conflict of interest is assumed between the organization and its members, and therefore the agents will simply maximize the organization's objective.

The analysis of the model yields three main results. First, a simple log-supermodularity condition is provided which ensures that, in balanced hierarchies, coordination time is increasing and the span of control is decreasing as one travels up the hierarchy. When this condition is fulfilled, senior managers spend more time planning and coordinating than junior managers and supervise fewer (direct) subordinates than their lower-level counterparts. Loosely speaking, the log-supermodularity condition requires coordination and the span of control not to be "too complementary" (in relative terms) in reducing duplications. Indeed, if the span of control has to be larger near the bottom of the organization, then it cannot be that the incentives to coordinate increase too sharply with the number of subordinates, or otherwise junior managers would spend more time planning and communicating than their superiors.

A second result is that coordination times and spans of control will typically be equalized across layers when concerns for fast execution are paramount. Thus, if senior managers coordinate more than junior managers (as the previous result suggests), then a shift toward granting junior

[2] In reality, however, not all duplications need to be wasteful, and indeed many authors have argued that redundancies can sometimes be beneficial when reliability is an issue. See Ting (2003), for instance.

managers broader authority in deciding how work is organized should be expected as urgency increases. Intuitively, we show that whereas communication overloads at the top levels of the hierarchy help reduce wage costs, bottlenecks are also created, and these are very costly when decisions need to be taken quickly. Thus, as urgency increases, a more decentralized allocation of responsibilities will often emerge.[3]

A third result relates to the optimal number of layers in a hierarchy. An unappealing feature of early information-processing models is that because delegation is only driven by the need to reduce delay, the number of levels in a hierarchy tends to increase with urgency (see, e.g., Keren and Levhari, 1979). In a simplified version of the model with uniform span of control across layers, we show that if returns to specialization are substantial, that result can be reversed. In particular, sufficient conditions are provided for the span of control to widen and the hierarchy to flatten as urgency increases. Empirically, the model suggests that if the main concern of an organization is to coordinate a specialized workforce cost efficiently, relatively tall hierarchies are optimal. However, a transition toward broader and flatter organizations should be observed as competitive time pressure grows.

These results may shed some light on dispersed observations about organizations. Many commentators have argued that managerial jobs entail a continuing effort to coordinate activities within the organization, and that planning and coordinating receive the greatest emphasis within top management (see, e.g., Sayles, 1964; Mahoney, Jerdee, and Carroll, 1965; Guetzkow, 1981). Starbuck (1971) also claims that "the span of control was supposed to be smaller near the top of the management hierarchy than near the bottom, because there was greater need for coordination near the top." Galbraith (1977) provides some support for that view by showing that in the production departments of U.S. and Canadian oil refineries, the span decreases as one passes from the second level (foreman) to the third (general foreman), and then again from the third to the fourth (superintendent). The present framework is consistent with these observations, provided of course that the log-supermodularity condition holds.[4]

In recent years, a trend toward the empowerment of lower-level managers and the decentralization and flattening of the firm has also been documented. Rajan and Wulf (2006) show that in the last 20 years, U.S. corporations have experienced a reduction in the number of formal layers and an increase in the number of managers directly reporting to the top management, and find a positive correlation between delayering and empowerment. Building on that work, Guadalupe and Wulf (2008) study the effect of trade liberalization between the United States and Canada on various measures of organization design. Their results suggest that greater international competition leads to flatter and broader firms. Similarly, Acemoglu et al. (2007) document a robust positive correlation between product market competition (measured as the inverse of the Lerner index) and various measures of decentralization in three data sets of French and British firms. Mendelson (2000) uses data collected from 63 business units in the information technology (IT) industry to test the idea that the choice of organizational form depends on the dynamism ("clockspeed") of the environment. He defines a measure of organizational IQ which combines several specific variables measuring the degree to which decision rights are decentralized, the importance of incentive pay, the adoption of focus strategies, and the like. He shows that this measure is strongly positively associated with profitability and growth and, more importantly, that organizational IQ has a stronger effect in the faster-moving segments of the IT industry than in the slower-moving ones. The latter result is interpreted as evidence that the adoption of a more

[3] We also consider a variant of the model where managers spend the same amount of time coordinating but differ in terms of their communicative skills. The main result there is that when urgency is paramount, not only should managers have similar communicative skills across layers, but these skills should also be as good as possible. This suggests that firms operating in turbulent environments should hire more talented managers than firms operating in stable environments, especially at the lower levels of the hierarchy.

[4] It is interesting to note that in the military, the span of control tends to be uniform across levels. This is also consistent with the model if, as Keren and Levhari (1979) argue, in the military the cost of staffing the hierarchy is secondary compared to the utility of planning time saved.

"organic" form of organization is driven by a fast clockspeed. Overall, therefore, the evidence seems consistent with the view that concerns for fast execution (or competitive time pressure) may be an important determinant of the recent trends toward decentralization and delayering in firms.

□ **Related literature.** This article contributes to a literature that studies the internal organization of firms from a viewpoint that abstracts from incentives. We share with the information-processing literature the view that, whereas individuals are limited in their ability to process information, organizations can help them overcome these limitations (see Van Zandt, 1998b, for an excellent survey). The seminal paper in this tradition is Radner (1993). Radner develops an explicit model of information processing based on the assumption that agents cannot process information instantaneously and shows that decentralizing operations through the use of a hierarchy is valuable to reduce delay. Bolton and Dewatripont (1994) also emphasize the benefits of hierarchical communication, but in a setting where the goal is to exploit returns to specialization (this is the second motive for delegation in the present article). Van Zandt (1998a) and Orbay (2002) extend Radner's framework to situations where new information arrives periodically, whereas Van Zandt (1999) considers scenarios where information arrives in real time. In all these models, information is hierarchically aggregated by boundedly rational agents. By contrast, the focus of this article is on how tasks should be subdivided when coordination problems are present. In particular, we allow for coordination activities (not only information processing) to take time, and study how such activities should be allocated across hierarchical levels.[5]

Keren and Levhari (1979) present a precursor to the information-processing literature which is closely related to the present work. They also study how concerns for fast execution affect the design of organizations and obtain results about the span of control that can be seen as a special case of ours. However, because there are no coordination activities in their model, they cannot highlight the role of complementarities between coordination and the span of control or study the decentralization of coordination responsibilities. Furthermore, greater urgency always leads to taller and narrower hierarchies in their model, whereas the opposite obtains here when returns to specialization are substantial. On a more technical side, this article differs from theirs (but also Keren and Levhari, 1989; Qian, 1994) as we do not rely on continuous approximations of the span of control or the number of levels in the study of hierarchies (we do, however, assume that the task can be infinitely subdivided). This is important because, as Van Zandt (1995) has shown, such approximations can be inaccurate, especially regarding the number of levels.

More recently, Meagher, Orbay, and Van Zandt (2003) and Van Zandt (2003) have examined how hierarchies evolve in a changing environment. In their setting, delay is costly because decisions based on old information become less appropriate over time. They find that the size of the organization's information-processing task is not invariant to conditions in its environment. Meagher et al., in particular, show that there is often a nonmonotonic, inverted-U relationship between the size of the task (and hence managerial size) and the volatility of the environment. In a resource-allocation model, Van Zandt finds that when the environment changes rapidly (as measured by the inverse of the correlation between old and new information) or when the environment is less volatile, small organizations tend to be optimal. Neither Meagher et al. nor Van Zandt emphasize, as we do, the interplay between specialization and delay in shaping organizations.

Finally, there are several papers that study hierarchies where either effort or ability is endogenous. Prat (1997) extends Radner's model to situations where agents differ in their ability to process information and shows that, under some conditions on the wage distribution, ability is uniform within layers and increasing with rank. In an efficiency wage model, Qian (1994) finds that effort expended on monitoring increases with rank in the hierarchy whereas the span

[5] In Vayanos (2003), hierarchical communication involves a loss of useful information; however, managers cannot exert effort to reduce such losses.

of control may increase or decrease depending on the specific utility function of the agents. Garicano (2000) focuses on the incentives to acquire knowledge in organizations. His model yields several predictions concerning the effects of changes in the costs of acquiring and transmitting information, as well as changes in the complexity of the production process. None of these papers studies how greater urgency affects organization design.[6]

The remainder of the article is organized as follows. Section 2 develops the model. Section 3 analyzes the simple case where the span of control is uniform across layers. Balanced hierarchies are studied in Section 4, where a variant of the model focusing on communicative skills is also presented. Section 5 focuses on delayering. Section 6 discusses the issues that arise when the size of the task is endogenous or new tasks arrive over time. Section 7 concludes. All the proofs are gathered in the Appendix.

2. The model

■ We consider an organization that must process a task of size $N \in \mathbb{R}_{++}$. There are two types of agents or "roles" in this organization: information processors ("workers") and coordinators ("managers"). Whereas workers are located at the bottom of the hierarchy and engage in production, managers are located at the upper levels-where they plan and coordinate the work of their direct subordinates.

Coordination is useful to reduce wasteful duplications of tasks that may arise during the division of labor. Consider a manager who delegates his task of size $M \leq N$ to s subordinates (called a work group). As in Becker and Murphy (1992), tasks can be subdivided into infinitely many subtasks. We assume that tasks are divided evenly among subordinates and that duplications may result during the division of labor. Specifically, we posit that the actual total amount of information processed by the subordinates is not M but $MD(s, t) \geq M$, where $s = 1, \ldots, \bar{s}$ is the span of control of the manager and $t \in [\underline{t}, \bar{t}]$ is the time that the manager spends coordinating his subordinates.[7] Thus, $D(s, t)$ measures lack of coordination or loss of control within the group.

The duplication function $D : \{1, 2, \ldots, \bar{s}\} \times [\underline{t}, \bar{t}] \to \mathbb{R}$ is characterized by the following three properties: (i) $D(s, t)$ is twice continuously differentiable in $t \in (\underline{t}, \bar{t})$, (ii) $D(s, t) \geq 1$, and (iii) $D_t(s, t) < 0$ for all $s \geq 2$.[8] These conditions impose mild restrictions on D. (i) is a technical assumption that is convenient to state the results of the article but could be easily dispensed with. (ii) states that we are dealing with duplications. (iii) captures the idea that duplications can be reduced by spending more time on planning and coordination activities. For now, we do not specify how coordination time and the span of control interact within the duplication function. However, some of the key results of the article will require $D(s, t)$ to be log-supermodular, which implies that the incentives to coordinate do not become too strong in relative terms as the number of subordinates grows. Note also that to ensure that the optimal span of control is always bounded, we assume that there is an upper bound $\bar{s} \geq 2$ to the number of subordinates that can be effectively supervised by a single manager. Similarly, we posit that there is a minimum amount of time $\underline{t} > 0$ that managers must spend coordinating for their work to be effective. This guarantees that hierarchies with an infinite number of layers are never optimal.[9]

Lastly, we describe how information is processed. Let w be the amount of information that a worker must process (his workload). We define a mapping $\eta : \mathbb{R}_{++} \to [\underline{\eta}, \bar{\eta}]$ and interpret $\eta(w)$

[6] On effort and ability in hierarchies, see also Williamson (1967), Calvo and Wellisz (1978), Keren and Levhari (1989), Geanakoplos and Milgrom (1991), and Meagher (2003), among others. Many of these models share with the present framework a recursive managerial production function and exhibit a tradeoff between a longer chain of command and a larger span of control.

[7] The time and effort it takes subordinates to absorb these instructions is assumed to be zero. This is similar to the convention adopted in the information-processing literature that all the costs of reporting are borne by superiors in the form of reading time.

[8] By convention, subscripts in functions of more than one variable will be used to denote partial derivatives.

[9] The last two assumptions could easily be relaxed, for instance by requiring $D(s, t)$ to be "large" whenever $s > \bar{s}$ or $t < \underline{t}$.

as the time it takes to process one unit of information when a worker's workload is w. Thus, $\eta(w)w$ is the time it takes to process w. The function η is assumed to be (weakly) increasing: $\eta' \geq 0$. This captures returns to specialization in information processing à la Becker and Murphy (1992), or overload costs. The idea is that a worker who concentrates on a narrow set of tasks is more specialized, and thus more productive, than a jack of all trades.[10]

□ **The problem of the hierarchy.** For any given amount of information N to be processed, the objective of the organization is to minimize total costs, which include not only the wages that must be paid to managers and workers but also the cost of delay. To fix ideas, let us begin with the simplest case of an organization composed of a single agent. In that case, the agent's workload is N and $\eta(N)N$ units of time are needed to process that information. The wage bill is thus $\nu\eta(N)N$, where ν denotes the wage per unit of time. With no loss of generality, this wage is normalized to unity throughout the rest of the article. Delay is the time the organization takes to process all the information, in this case $\eta(N)N$. The cost of delay is $C(\eta(N)N)$, where $C(\cdot)$ is a strictly increasing function. The total cost associated with a single-agent organization is thus

$$\eta(N)N + \lambda C(\eta(N)N), \tag{1}$$

where $\lambda > 0$ denotes the weight attached to the cost of delay relative to wage costs.

Now consider a two-layer hierarchy. Suppose the top manager at layer 1 delegates the task to a number of subordinates (say s_1) at layer 2, who then process the information. Because of duplications, the size of the task at the second level is not N but $ND(s_1, t_1)$, where t_1 denotes the time the top manager spends on coordinating. If tasks are divided evenly among subordinates, individual workloads at layer 2 are given by $w_2 = ND(s_1, t_1)/s_1$. Therefore $\eta(w_2)w_2$ is the time each worker must spend processing information. Total costs are thus given by

$$\eta(w_2)ND(s_1, t_1) + t_1 + \lambda C(\eta(w_2)w_2 + t_1). \tag{2}$$

The comparison between (1) and (2) highlights some of the key costs and benefits of delegation. On the positive side, delegation to multiple subordinates allows the organization to expand its size and more finely subdivide labor. A finer division of labor in turn benefits the organization because it increases specialization (because in the optimum, $w_2 < N$) and tends to reduce delay through parallel information processing. Note in fact that delay includes only the time spent processing information by *one* worker at layer 2 because all the agents at the same level are assumed to work concurrently. (It should also be clear that because none of these benefits accrue when $s = 1$, delegation to a single subordinate cannot be optimal; see Proposition 1 below.) Turning to costs of delegation, these include the wages that managers must be paid for doing their job and the delay that coordination activities such as meetings and deliberations bring about. Moreover, because coordination will in general be imperfect, the total amount of information to be processed will increase (from N to $ND(s_1, t_1)$ in this case), with negative repercussions on both labor costs and delay.

Next, consider a three-layer hierarchy. As before, let s_1 be the top manager's span of control and t_1 the time he spends coordinating his subordinates. Then each manager at layer 2 is in charge of coordinating a task of size $ND(s_1, t_1)/s_1$. Suppose that each delegates his task to s_2 subordinates and spends t_2 units of time coordinating their work. A worker's workload at layer 3 is thus $w_3 = ND(s_1, t_1)D(s_2, t_2)/s_1s_2$. Because there are s_1s_2 workers, total processing time is $\eta(w_3)ND(s_1, t_1)D(s_2, t_2)$, and the total cost associated with a three-layer hierarchy is therefore

$$\eta(w_3)ND(s_1, t_1)D(s_2, t_2) + t_1 + s_1t_2 + \lambda C(\eta(w_3)w_3 + t_1 + t_2). \tag{3}$$

The three-layer example neatly illustrates how coordination responsibilities are delegated across managerial layers. Whereas coordination *among* the work groups at layer 3 is ensured by the effort

[10] Bolton and Dewatripont (1994) also stress the importance of specialization in information processing. In their model, however, gains from specialization arise from the repetition of the same task.

of the top manager, coordination *within* those work groups is provided by the middle management. Top and middle managers thus jointly ensure that the activities of the organization are properly coordinated. This example also clarifies the role of the assumption that tasks are divided evenly among subordinates. In fact, if all managers at the same level face exactly the same problem, the optimal values of s and t at that level will be the same and the resulting hierarchy will be balanced.[11]

We now tackle the general case of arbitrary balanced hierarchies. A hierarchy is defined as a rooted tree in which an agent's level or layer is given by the maximum distance from the agent to one of the workers inferior to the agent (for a formal definition, see Van Zandt, 1998b). A hierarchy is said to be balanced if it has no skip-level reporting (i.e., the path from the top manager to each worker has the same length) and all the managers at the same level have the same number of subordinates and spend the same time coordinating. Thus, a balanced hierarchy is characterized by a triple $\langle \mathbf{s}, \mathbf{t}, L \rangle$, where $L \geq 2$ is the number of levels in the organization (indexed $l = 1, \ldots, L$ from top to bottom), $\mathbf{t} = (t_1, \ldots, t_{L-1})$ is a vector of per-manager coordination times at each level of the hierarchy, and $\mathbf{s} = (s_1, \ldots, s_{L-1})$ is a vector of spans of control at each level of the hierarchy.

The cost associated with hierarchy $\langle \mathbf{s}, \mathbf{t}, L \rangle$ is obtained as follows. Let the number of managers at layer l be given by $n_l \equiv \prod_{k=0}^{l-1} s_k$ (by convention, $n_1 \equiv s_0 \equiv 1$ refers to the top manager). The process of sequential delegation of tasks illustrated above yields the following expression for a worker's workload at layer L:

$$w_L = N \prod_{l=1}^{L-1} \frac{D(s_l, t_l)}{s_l}.$$

Total information-processing time is thus $\eta(w_L) N \prod_{l=1}^{L-1} D(s_l, t_l)$. Total working time (information processing plus coordination) and hence the wage bill are given by

$$\eta(w_L) N \prod_{l=1}^{L-1} D(s_l, t_l) + \sum_{l=1}^{L-1} n_l t_l. \tag{4}$$

A second important criterion for evaluating the effectiveness of an organization is how swiftly it implements managerial directives. The time it takes the hierarchy to process all the information is given by

$$\eta(w_L) w_L + \sum_{l=1}^{L-1} t_l. \tag{5}$$

Recall in fact that only one employee at each level of the hierarchy must be considered when computing total delay because managers at the same layer are assumed to work in parallel.

The problem of the hierarchy is to minimize a weighted sum of the wage bill and the cost of delay with respect to the number of layers in the hierarchy L, the spans of control $\mathbf{s} = (s_1, \ldots, s_{L-1})$, and the coordination times $\mathbf{t} = (t_1, \ldots, t_{L-1})$:

$$\min_{\mathbf{s},\mathbf{t},L} \eta(w_L) N \prod_{l=1}^{L-1} D(s_l, t_l) + \sum_{l=1}^{L-1} n_l t_l + \lambda C \left(\eta(w_L) w_L + \sum_{l=1}^{L-1} t_l \right). \tag{6}$$

[11] At this point it may be helpful to compare the present framework, and in particular (3), with Keren and Levhari's (1979) seminal model. Using the notation already introduced, the cost associated with a three-layer hierarchy in Keren and Levhari is given by $1 + s_1 + \lambda C(a(s_1 + s_2) + 2b)$, where a and b are positive constants. $1 + s_1$ is the total number of managers in the hierarchy, who are paid a fixed wage ($1 + s_1$ is therefore also the wage bill). $s_1 s_2$ is exogenously given and measures the number of shops (or workers) in the firm. The time it takes a manager to process his information is linear in his span of control and is given by $as + b$. Note that there are no coordination activities or specialization in their model. Furthermore, unlike them, we allow for an endogenous number of workers. Finally, Keren and Levhari's assumption that managers are salaried is problematic because the time managers spend working is endogenous. In our model, by contrast, all agents are paid by the hour.

Delay is assumed to be costly ($C' > 0$), and λ measures the weight attached to the cost of delay relative to the wage bill. We will also sometimes write $\lambda = +\infty$ to denote a situation when only delay matters. This should be interpreted as saying that the weight attached to the wage bill, v, is zero. Of course, if $\lambda = +\infty$, the problem of the hierarchy reduces to minimizing (5) with respect to $\mathbf{s}$, $\mathbf{t}$, and L.

Proposition 1. A solution to the problem of the hierarchy exists. Moreover, in the optimum, every manager must have at least two subordinates: $s_l \geq 2 \; \forall l = 1, \ldots, L-1$.

The proof of Proposition 1 is omitted but the key steps are as follows. The fact that $\underline{t}$ and $\bar{s}$ are bounded and strictly positive implies that the size $(\mathbf{s}, L)$ of an optimal hierarchy must be finite. Furthermore, for any given $(\mathbf{s}, L)$, a solution to the problem of the hierarchy with respect to $\mathbf{t}$ exists by the Weierstrass theorem. Then, because one can always choose a maximum element from a finite set, there must exist a solution to (6) and to minimizing (5). Note also that delegation to a single subordinate is never optimal. If $s_l = 1$, in fact, no gains from specialization or reduction in delay accrue at layer l. The organization would therefore be better off if the manager at layer $l-1$ did not delegate his task, thus saving at least the coordination cost $\underline{t}$.

3. Optimal coordination in uniform hierarchies

■ To develop some intuition for the forthcoming results, this section focuses on a simple scenario where the span of control is the same across layers and equal to s (a uniform hierarchy). The analysis of the general case is postponed to the next section.

Let $n_l = s^{l-1}$ and $w_L = N(\prod_{l=1}^{L-1} D(s, t_l))/s^{L-1}$. For any given $L \geq 2$ and $s \geq 2$, the problem of the uniform hierarchy is to minimize

$$\eta(w_L) N \prod_{l=1}^{L-1} D(s, t_l) + \sum_{l=1}^{L-1} n_l t_l + \lambda C \left(\eta(w_L) w_L + \sum_{l=1}^{L-1} t_l \right) \tag{7}$$

with respect to $\mathbf{t} = (t_1, \ldots, t_{L-1})$. The following proposition characterizes the optimal assignment of coordination responsibilities in this case.

Proposition 2.

(i) Suppose wage costs are not negligible ($\lambda < +\infty$). Then, in the optimum, higher-level managers spend more time coordinating than lower-level managers: $t_1 \geq \cdots \geq t_{L-1}$.

(ii) If instead only delay matters and coordination is more beneficial in percentage terms when it is scarce (i.e., $D_t(s, t)/D(s, t)$ is strictly increasing in t),[12] then in the optimum, coordination times are equalized across layers: $t_1 = \cdots = t_{L-1}$.[13]

Proposition 2 implies that in the optimum, senior managers work more and are therefore paid more than junior managers. Loss of control also becomes more severe as one goes down the hierarchy because $D(s, t_l) \leq D(s, t_{l+1})$ for all l and s. The intuition is simple: in the optimum, higher-level managers coordinate more than lower-level managers because the former exert their influence on a much greater portion of the hierarchy. Consider the wage bill in isolation. Although the various duplication functions (and hence the coordination times) enter into the wage bill symmetrically, the individual (per-capita) contribution of managers at different layers is different. Indeed, there are fewer and fewer managers as one goes up the hierarchy. Thus, their

[12] In this article, we adopt the convention that the term "increasing" stands for "nondecreasing" (or "weakly increasing"). When as in Proposition 2 we want an inequality to be strict, we will say so explicitly, as in "strictly increasing", and so forth.

[13] There are actually several assumptions that ensure that the solution to this minimization problem is such that $t_1 = \cdots = t_{L-1}$. For instance, uniqueness, and hence the result, would follow from the strict convexity of delay in $\mathbf{t}$. Less obviously, one could also show that uniqueness is implied by the strict submodularity of delay (I thank an anonymous referee for pointing this out to me).

actions have a much greater impact on the performance of the organization than those taken by their subordinates. By contrast, when computing delay, the efforts of individuals at different layers of the hierarchy enter (7) symmetrically because only the contribution of a single manager at each level of the hierarchy has to be considered. This explains why, when only delay matters, coordination times are typically equalized across layers. Because the optimal level of coordination results from a combination of these two effects, the introduction of delay tends to weaken the wage-cost effect but cannot cancel it out, unless of course only delay matters.

Clearly, Proposition 2 provides a rationale for the view that the time a manager spends in planning and coordination activities generally increases with rank in the hierarchy, especially when the difference in level between managers is high. Proposition 2 also suggests a possible explanation for the recent trend toward decentralization in firms. According to the model, in fact, a shift toward granting employees broader authority in deciding how work is coordinated should be expected when urgency increases. Together, these results can be interpreted as formalizing the presence and indeed the optimality of communication overloads at the top levels of the hierarchy, overloads which, however, are suboptimal from a purely delay-minimizing point of view.

We conclude this section with an example that illustrates Proposition 2 and shows that, as urgency increases, differences in coordination times across layers may decrease monotonically, not just disappear in the limit. (Proofs of the examples are also provided in the Appendix.)

Example 1. Let $D(s, t) = s^{P(t)}$ where $P(t) = 1 - t^{\alpha}$, $\alpha \in (0, 1)$, $\bar{t} \leq 1$. Moreover, let $C(\cdot)$ be linear in delay and $\eta(\cdot) = \eta$. Then the problem of the uniform hierarchy is strictly convex in $\mathbf{t}$ and, assuming interior solutions, in the optimum we have $t_1 > \cdots > t_{L-1}$, $\partial(t_l/t_{l+1})/\partial\lambda < 0$, and $\lim_{\lambda \to +\infty} t_l/t_{l+1} = 1$ for all $l \leq L - 2$.

4. Balanced hierarchies

■ In this section, we tackle the general case where the span of control can also change across layers. This problem is considerably more difficult than the one in the previous section because many more endogenous variables are involved, some of which can only take integer values. Nevertheless, a partial characterization of the solution can be provided.

Proposition 3. Consider the problem of the hierarchy in (6). Then, in the optimum, either senior managers coordinate more than junior managers or have a smaller span of control. That is, for all i, j with $i < j$, either we have $t_i > t_j$ or $s_i < s_j$, unless $s_i = s_j$ and $t_i = t_j$.

Proposition 3 provides some support for the view that managers should coordinate more and have smaller spans of control near the top of the hierarchy. Small spans of control near the top of the hierarchy tend to be beneficial because they allow the organization to minimize the number of people that must be employed in managerial positions (keeping the number of workers fixed), especially near the top where coordination requirements are substantial. Technically, the proof exploits the fact that, so long as $t_i < t_j$ and $s_i \geq s_j$ (or $t_i \leq t_j$ and $s_i > s_j$), one can always reduce the total wage cost of coordination $\sum_{l=1}^{L-1} n_l t_l$ while keeping both delay and the workers' information-processing workload constant by swapping (s_i, t_i) with (s_j, t_j). However, the result falls short of showing that both small spans of control and longer hours are always optimal near the top of the organization. Indeed, to do that, one must specify how coordination time and span of control interact to reduce duplications.

Proposition 4. Suppose $D(s, t)$ is log-supermodular.[14] Then, in the optimum, senior managers coordinate more than junior managers and have smaller spans of control: $t_1 \geq \cdots \geq t_{L-1}$ and $s_1 \leq \cdots \leq s_{L-1}$.

[14] Log-supermodularity of a positive function $h(x, y)$ implies that the relative returns, $h(x_H, y)/h(x_L, y)$, are increasing in y for all $x_H > x_L$. Obviously, this condition is only needed here to hold on the relevant domain $\{(s, t) \in \mathbb{N} \times \mathbb{R} : 2 \leq s \leq \bar{s}, \underline{t} \leq t \leq \bar{t}\}$. For formal definitions of the monotone comparative statics concepts used in this article, see Topkis (1998).

Proposition 4 highlights the key role of complementarity between coordination and span of control. The log-supermodularity of $D(s, t)$ essentially requires coordination not to reduce duplications too much when the span of control is large relative to when it is small. Intuitively, if the span of control has to be larger near the bottom of the organization, then it cannot be that the incentives to coordinate increase very sharply (in relative terms) with the number of subordinates, or otherwise junior managers might end up working more than their superiors. Put differently, the log-supermodularity condition places an upper bound on the strength of the complementarity between s and t which is sufficient for the result to hold.

Examples of duplication functions which are log-supermodular can easily be constructed. For instance, let $D(s, t) = s^{H(s,t)}$, where $H(s, t) = 1 - (\frac{t}{s})^{\sigma}$ if $1 \geq (\frac{t}{s})^{\sigma}$ and 0 otherwise, for $\sigma > 0$ and $s \in [2, \bar{s}]$. Clearly there are no duplications here if a manager spends at least one unit of time coordinating each subordinate, and attention can be restricted to the case where $1 \geq (\frac{t}{s})^{\sigma}$. One can easily check that D is log-supermodular if $\sigma \geq 1/\ln(2)$. Furthermore, if $\sigma = 1$, D is log-supermodular provided $s \geq 3$.

It is also important to stress that the log-supermodularity of D does not imply that, in absolute terms, the incentives to coordinate decrease with group size (that is, D need not be supermodular). To see that, suppose $D(s, t) = v(s)r(t)$. This function is clearly log-supermodular; however, if v is increasing in s, then D is also submodular.

Lastly, we emphasize that the log-supermodularity of D is only a sufficient condition, which becomes less and less stringent as the hierarchical gap between two managers grows. Note in fact that, so far as the wage bill is concerned, the marginal cost of coordination grows exponentially as one goes down the hierarchy because $n_l = \prod_{k=0}^{l-1} s_k \geq 2^{l-1}$. Thus, if the organization cares about minimizing wage costs and the hierarchical gap is large, it is likely that senior managers will coordinate more than junior managers, no matter what their respective spans of control are.[15]

We now turn to the case when only delay matters. From the fact that (5) is invariant with respect to permutations of pairs (s_l, t_l) in $(\mathbf{s}, \mathbf{t})$, the following result immediately obtains.

Proposition 5. Suppose that the problem of the hierarchy when only delay matters has a unique solution. Then, in the optimum, coordination times and spans of control are equalized across layers: $t_1 = \cdots = t_{L-1}$ and $s_1 = \cdots = s_{L-1}$.[16]

Together, Propositions 4 and 5 suggest that senior managers will often spend more time coordinating and have smaller spans of control than junior managers, but also that these differences should disappear as λ approaches infinity. Importantly, this holds true despite the fact that the optimal number of layers in the hierarchy will in general change as λ varies, because our propositions hold for all L, not necessarily the same. In the next section, we will study how the shape of the hierarchy changes as urgency increases. Before doing that, however, we briefly discuss a variant of the model where managers differ in their ability to coordinate.

□ **Communicative skills.** An implication of the above analysis is that higher-level managers typically spend more time on coordination activities than lower-level managers simply due to their different positions in the hierarchy, not because of inherent differences in ability. Indeed, because managers are identical, it does not really matter which hierarchical position a manager is allocated to. Casual empiricism, however, suggests that the ability to coordinate effectively is very important, especially at the top of the hierarchy. It is in fact often required that successful

[15] A simple example may help clarify the magnitude of this effect. Suppose that $n_1 = 6$ and $n_l = 8$, $l = 2, \ldots, 6$. Then, the cost in terms of wages of increasing coordination time is almost 200,000 times bigger at layer 6 than at layer 1, because there are many more managers at layer 6. Thus, it is likely that in the optimum, t_1 will be greater than t_6, even if coordination time is more effective at reducing duplications when the span of control is large.

[16] Due to integer constraints, it is difficult to minimize (5) with respect to $\mathbf{s}$ and $\mathbf{t}$ and to show that a solution is unique. However, ignoring such constraints, one can assume that the program is strictly convex. The true solution will then be one of the integer solutions nearest to the unique solution found and, if the organization is large, the relative error will be small.

candidates for executive positions have "strong interpersonal and communicative skills, a collegial management style," "ability to communicate effectively," and so forth.[17]

This subsection considers a variant of the model where managers differ in their communicative skills (or "ability"). The basic idea is that coordination problems can be mitigated by hiring better communicators. To keep things simple, suppose that all the managers spend the same amount of time $\bar{\tau}$ coordinating the work of their subordinates. The main changes are that t now denotes communicative skills and managerial wages depend on ability. Specifically, let $\omega(t)$ be the wage paid to a manager of ability t working $\bar{\tau}$ hours. Clearly $\omega' > 0$ because higher-ability managers must be paid more than their lower-ability counterparts. The other features of the model remain the same, with the understanding of course that $D(s, t)$ now measures the duplications that occur when a manager has s subordinates and his ability is t. The problem of the hierarchy when managers differ in their communicative skills is then to minimize

$$\eta(w_L)N\prod_{l=1}^{L-1} D(s_l, t_l) + \sum_{l=1}^{L-1} n_l\omega(t_l) + \lambda C(\eta(w_L)w_L + (L-1)\bar{\tau}) \tag{8}$$

with respect to $\mathbf{s}$, $\mathbf{t}$, and L. The following result, whose proof is analogous to that of Proposition 4 and is thus omitted, characterizes the solution to this problem.

Proposition 6. Consider the problem of the hierarchy in (8) and suppose $D(s, t)$ is log-supermodular. Then in the optimum $t_1 \geq \cdots \geq t_{L-1}$ and $s_1 \leq \cdots \leq s_{L-1}$. Furthermore, when only delay matters, $t_1 = \cdots = t_{L-1} = \bar{t}$, regardless of whether or not $D(s, t)$ is log-supermodular.

The first part of Proposition 6 is consistent with the view that managers with strong interpersonal and communicative skills are typically found in the top echelons of the hierarchy. This is not surprising, and similar results have been obtained for instance by Geanakoplos and Milgrom (1991) and Prat (1997). More interesting is the fact that when urgency is paramount, not only are ability levels equalized across layers but also they are as high as possible. Note in fact that whereas in the main model, coordination is costly both in terms of wages and delay, here ability only adversely affects (8) because higher-ability managers command higher wages. Thus, only the ablest individuals will be employed by the organization when delay is all that matters.

Empirically, this result suggests that firms operating in turbulent environments (and for which delay is presumably very costly) should hire managers of higher ability and provide more training opportunities to existing employees than companies operating in mature sectors. Furthermore, these differences should be more pronounced at the lower and middle levels of the hierarchy, where minimizing wage costs is especially important.

5. The flattening hierarchy

■ We now turn to the issue of how hierarchies evolve as concerns for fast execution become more important. The analysis is motivated by evidence that in the last 20 years, U.S. corporations have become flatter while the number of managers directly reporting to the top management has increased (see Rajan and Wulf, 2006). Our contribution will be to show that, if returns to specialization are substantial, both trends can be rationalized as an optimal response to increased time pressure.

In light of the empirical motivation, we consider a simple version of the model focusing on the (average) span of control s and the number of levels in the hierarchy L. Specifically, suppose that all the managers spend a given amount of time $\bar{\tau}$ coordinating the work of their subordinates and that the span of control is uniform across layers. Then the hierarchy is fully characterized by s and L. We adopt the shorthand notation $d(s) \equiv D(s, \bar{\tau})$ and posit for simplicity that

[17] Survey evidence also indicates that when recruiting new production staff, U.S. employers rank communicative skills above previous work experience, recommendations, years of schooling and grades, and so on. See Bureau of the Census (1995).

$d(s+1) \geq d(s)$ and $\frac{d(s+1)}{s+1} \leq \frac{d(s)}{s}$ for all $s = 1, \ldots, \bar{s}-1$. Lastly, we posit that, whenever a task is delegated and a new work group is formed, a delay $b \geq 0$ occurs, which is not accounted for by the time $\bar{\tau}$ managers spend coordinating. This delay has no associated wage bill and is meant to capture a number of factors that often lengthen the decision-making process, most notably the difficulty of getting all the relevant parties together.[18] In the model, the introduction of b helps derive a simple, sufficient condition in Example 2 below; however, it plays no role in the derivation of the comparative statics result of this section.

Let $n_l = s^{l-1}$ and $w_L = N(d(s)/s)^{L-1}$. The problem of the hierarchy is now to minimize

$$T(s, L; \lambda) = WB(s, L) + \lambda C(DL(s, L)) \tag{9}$$

with respect to s and L, where

$$WB(s, L) = \eta(w_L) N d(s)^{L-1} + \sum_{l=1}^{L-1} n_l \bar{\tau} \tag{10}$$

and

$$DL(s, L) = \eta(w_L) w_L + (L-1)(\bar{\tau} + b). \tag{11}$$

As before, attention can be restricted to the case where $s \in [2, \bar{s}]$ and, for simplicity, we rule out one-person organizations ($L = 1$). Note that, because $d(s)/s$ is decreasing in s, when λ is large the optimal span of control will be close to the maximum number of subordinates $\bar{s}$ that can be efficiently supervised.[19] Thus, the model naturally captures the idea that broad organizations tend to be optimal when concerns for fast execution are important. Sharp predictions, however, can only be obtained if more structure is imposed to the model. We begin with a set of assumptions that implies that s and L are substitutes in the problem of the hierarchy.

Assumption (S). (i) $WB(s, L)$ and $DL(s, L)$ are supermodular in (s, L) on $S \equiv \{(s, L) \in \mathbb{N} \times \mathbb{N} : \bar{s} \geq s \geq 2, L \geq 2\}$. (ii) For fixed $s \in [2, \bar{s}]$, $WB(s, L)$ and $DL(s, L)$ are quasi-convex in L. (iii) $C'' \leq 0$.

That the span of control and the number of levels should be substitutes is quite intuitive: after all, s and L are alternative ways to increase the size of the organization and more finely subdivide tasks. Mathematically, that notion is captured by the requirement that $T(s, L; \lambda)$ be supermodular in (s, L). Unfortunately, the supermodularity of *WB* and *DL* does not necessarily imply that T is supermodular (or even quasi-supermodular), unless C is linear. This is why parts (ii) and (iii) of Assumption (S) are needed. Indeed, it can be shown that if the domain of s and L is opportunely restricted (thanks to the quasi-convexity of *WB* and *DL*), T will be supermodular provided C is concave or, more generally, not too convex (see the proof of Proposition 7).

Functional forms consistent with (S) are not hard to find. For instance, if $d(s) = \beta s^{\alpha}, \alpha < 1, \beta \in (2^{-\alpha}, 1)$, and $\eta(w) = \phi w^{1+\psi}, \psi \geq 0$, then *WB* and *DL* are quasi-convex in L and, provided ψ is large enough, also supermodular in (s, L) for all $s, L \geq 2$.[20] Note in particular that the convexity of η realistically implies decreasing returns to specialization in information processing, and ψ measures the strength of the returns to specialization.

In addition to (S), a second condition will be needed for our analysis. Loosely speaking, this condition states that the number of levels that minimizes wage costs is at least as large as the number of layers that minimizes delay. The condition is obtained as follows. Let

[18] Practitioners often emphasize this type of delay. For instance, Robert J. Herbold, former COO at Microsoft, reports that planning exercises are frequently delayed or cancelled because "key people aren't currently available" (Herbold, 2002). Moreover, if $\bar{\tau}$ is large, b would then also include the time managers need to rest and sleep.

[19] More generally, one might just require $\arg\min d(s)/s$ to be "large."

[20] Note that if $(1-\alpha)(1+\psi) > \alpha$, then both *WB* and *DL* are decreasing in s when $L = 2$, which implies that in the optimum there can be more than two levels in the hierarchy only if $\bar{s}$ is not "too large" (that is, if the span of control cannot be stretched indefinitely). Arguably, this last requirement on $\bar{s}$ is unlikely to be too restrictive, especially if N is large.

$S_L = \{L \in \mathbb{N} : L \geq 2\}$. For given $s \in [2, \bar{s}]$, let $L_W^+(s)$ be the greatest $\arg\min_{L \in S_L} WB(s, L)$ and $L_D^-(s)$ be the smallest $\arg\min_{L \in S_L} DL(s, L)$. We say that delegation is mainly driven by specialization, and write $L_W^+ \geq L_D^-$, if $L_W^+(s) \geq L_D^-(s)$ for all $s \in [2, \bar{s}]$. This terminology is motivated by the fact that L_W^+ can be large only if returns to specialization are substantial; further justification is provided below in Example 2.

We are now in a position to establish the following:

Proposition 7. Suppose Assumption (S) holds and delegation is mainly driven by specialization. Then as urgency increases, the hierarchy flattens and the span of control widens.

Proposition 7 suggests that concerns for fast execution may be key to explaining the evidence on the evolution of corporate hierarchies documented by Rajan and Wulf (2006). The proof is based on a simple idea. The key observation is that, if $L_W^+ \geq L_D^-$, the hierarchy tends to become flatter as urgency increases because the optimal number of levels in the organization is bigger than the number of levels that minimizes delay.[21] Furthermore, the span of control tends to increase with urgency because, so far as delay is concerned, larger spans are always beneficial.[22] These two effects reinforce each other because s and L are substitutes on the relevant domain. The key issue is of course when one can reasonably expect the condition $L_W^+ \geq L_D^-$ to hold. Intuition suggests that this condition is more likely to be fulfilled when returns to specialization are substantial, because L_W^+ will be large, and when getting all the relevant parties together is difficult (i.e., $b > 0$), because L_D^- will be small. The following example supports this conjecture:

Example 2. Suppose $d(s) = \beta s^\alpha$ and $\eta(w) = \phi w^{1+\psi}$, where $\alpha < 1$, $\beta \in (2^{-\alpha}, 1)$, and $\psi \geq 0$. Furthermore, suppose $(1 - \alpha)(1 + \psi) > \alpha$.[23] For every $b > 0$, there exists a $\bar{\psi}$ such that, if $\psi \geq \bar{\psi}$, then $L_W^+ \geq L_D^-$.

Proposition 7 and Example 2 provide a simple explanation for the recent move toward flatter hierarchies and broader spans of control. When the main concern of a firm is to coordinate its workforce cost efficiently and returns to specialization are substantial, relatively tall and narrow hierarchies are optimal. Such hierarchies, however, generate large delays because tasks must be delegated several times. Thus, as urgency grows, a transition toward flatter and wider organizations should be observed.

From an applied perspective, these findings suggest that the trend toward flatter organizations should be more accentuated in previously protected or slow-moving industries where returns to specialization are substantial. In that respect, it is interesting to note that some major reorganizations have taken place in industries such as microprocessors and oil extraction where gains from specialization are arguably very large (e.g., Intel in 2005 and British Petroleum in the 1990s). However, more empirical work is needed to test the predictions of this model.

6. Robustness

■ This section briefly examines the consequences of relaxing some of the assumptions of the article.

□ **The optimal size of the task.** We have so far assumed that the size of the information-processing task is exogenously given. This could be a valid approximation if, for instance, the organization is a contractor that must deliver a well-specified product. However, in many situations, it is more plausible to assume that the organization has some latitude over the characteristics of the product or service it provides. These considerations can easily be

[21] In the optimum, in fact, $L \in [L_D^-(s), L_W^+(s)]$ because WB and DL are quasi-convex.

[22] However, the assumption that $d(s)/s$ is decreasing in s (or more generally that $DL(s, L)$ is decreasing in s) is not essential for the result. An alternative set of assumptions is provided in the Appendix after the proof of Proposition 7.

[23] This condition ensures that the information-processing costs in (10) decrease both in s and L and therefore that there are genuine tradeoffs so far as the wage bill is concerned.

incorporated into the model. A standard procedure is to first find the least costly way to process N (this is the problem studied in this article) and then optimize over N, for a given benefit function $R(N)$.

A natural question is whether endogenizing the size of the task would alter the qualitative results of the article. Clearly the results concerning how coordination times and spans of control vary across levels would not be affected at all because they hold for any N. The comparative statics, however, would become more complex. Whereas in fact the effects of increased urgency on the structure of the hierarchy highlighted in the previous section would still be present even when N is endogenous, there would now be an effect operating from λ to N and then from N to (s, L). Of course, to the extent that the latter effect is small, our results would still go through. And finally, despite the complications due to these indirect effects, it would be interesting to consider briefly how an increase in urgency may affect the size of the task. There is in fact an obvious incentive for N to fall as urgency increases because delay rises with N. Thus, we would expect smaller, more-focused firms to emerge as competitive time pressure grows.[24]

□ **Periodic arrival of tasks.** This article posits that the organization must process a single task (a "one-shot" model). More realistically, one could envisage situations where new tasks arrive periodically. In the information-processing literature, Van Zandt (1998a) and Orbay (2002) have studied this type of model. Orbay, in particular, considers a setting where independent cohorts of data of the same size arrive at fixed intervals and tasks must be processed by the organization in the same way by the same managers (i.e., procedures are *stationary*).

Stationarity places additional constraints on hierarchy design. In fact, because information processing takes time, the organization can keep up with the arrival of new information only if the arrival rate is not too big. Following Orbay, let τ_c be the time interval between the arrival of successive tasks. In the main model, for stationarity to hold, the following throughput constraint must be met: $\tau_c \geq \max\{t_1, \ldots, t_{L-1}, \eta(w_L)w_L\}$. Thus, as Van Zandt (1998b) has noted, it is the agent with the largest workload who is effectively the bottleneck limiting the arrival of new tasks. Note also that the concept of urgency is enriched, being captured both by the cost of delay λ once a new task has arrived and by the frequency $1/\tau_c$ with which new tasks arrive. A simple way to incorporate the throughput constraint into the present model is to replace $\bar{t}$ with τ_c. Care should also be taken to insure that $\tau_c \geq \eta(w_L)w_L$, if necessary by increasing the spans of control or the number of levels in the hierarchy. But fortunately, provided that a stationary hierarchy exists,[25] the qualitative results of this article are not affected by these restrictions.

To see this, consider a stationary hierarchy $\langle \mathbf{s}, \mathbf{t}, L \rangle$ such that Proposition 3 does not hold. Then one can find another hierarchy $\langle \mathbf{s}', \mathbf{t}', L \rangle$ which also satisfies the throughput constraint but is associated with a strictly lower total cost (the rearrangement of $(\mathbf{s}, \mathbf{t})$ used to prove Proposition 3 would do, for instance). Similar remarks apply to Propositions 4 and 5. Thus, our characterization results extend to stationary hierarchies. Turning to the comparative statics on (s, L), the main complication there arises from the fact that additional constraints must be placed on s and L. Indeed, s and L must be sufficiently large so that the throughput constraint $\tau_c \geq \eta(w_L)w_L + b$ holds ($\tau_c \geq \bar{\tau} + b$ must also hold). These restrictions, however, can be incorporated into the proof of Proposition 7 without affecting its conclusion (the proof is available from the author upon request). Thus, provided a stationary hierarchy exists, the key qualitative results of this article remain valid.[26]

[24] For an analysis of these issues (in different environments), see Meagher, Orbay, and Van Zandt (2003) and Van Zandt (2003).

[25] τ_c may be so small that no stationary hierarchy exists. This would be the case, for instance, if $\tau_c < \underline{t}$ in the main model or $\tau_c < b$ in the model of Section 5. Existence of a stationary hierarchy thus requires τ_c to be larger than some minimum threshold.

[26] It is also important to note that in periodic models, some agents are typically idle some of the time. Idle time is not costly here because agents are assumed to be paid by the hour. However, if managers were paid fixed salaries, it would be important to minimize such idle time. See Van Zandt (1998a) for an analysis of the complications that would arise in that case.

7. Conclusion

■ Coordination is an essential ingredient for survival and success in competitive environments, and ways to enhance it are a central concern for modern management. Formal meetings, liaison groups, and teamwork are just a few examples of how firms try to promote exchange of information among interdependent organizational units. Using a model which explicitly recognizes the fact that communication takes time, this article studies how coordination responsibilities should be allocated across hierarchical levels and how the organization should evolve as concerns for fast decision making and execution become more important, relative to wage costs. Our result suggest that considerations of delay might have been an important determinant of the recent trends toward decentralization and delayering in firms. The analysis, however, has neglected many important issues, most notably the role of incentives. Studying the interrelations between coordination, incentives, and delay is an exciting avenue for future theoretical research.

Appendix

■ Proofs of Propositions 2, 3, 4 and 7 and Examples 1 and 2 follow.[27]

Proof of Proposition 2.

(i) We begin with the case where wage costs are not negligible. Let (7) be denoted by $T(\mathbf{t})$. Suppose $\mathbf{t}$ is such that $t_i < t_j$ for some $i < j$. Let $\mathbf{t}'$ be such that $t'_l = t_l$ for all $l \neq i, j$ and $t'_i = t_j$ and $t'_j = t_i$. Then $T(\mathbf{t}) - T(\mathbf{t}') = n^{i-1}t_i + n^{j-1}t_j - n^{i-1}t_j - n^{j-1}t_i = (n^{j-1} - n^{i-1})(t_j - t_i) > 0$. Thus, $\mathbf{t}$ is not optimal.

(ii) When only delay matters, the problem of the (uniform) hierarchy is to minimize

$$DL(\mathbf{t}) \equiv \eta(w_L(\mathbf{t}))w_L(\mathbf{t}) + \sum_{l=1}^{L-1} t_l \tag{A1}$$

with respect to $\mathbf{t}$. (Here individual workloads w_L are denoted by $w_L(\mathbf{t})$ to make explicit their dependency on $\mathbf{t}$.) The second part of Proposition 2 can be proven by showing that (a) if $w_L(\mathbf{t})$ is strictly convex, then in the optimum $t_1 = \cdots = t_{L-1}$, and that (b) if $D_t(s,t)/D(s,t)$ is strictly increasing in t, then $w_L(\mathbf{t})$ is strictly convex.

To prove (a), suppose $\mathbf{t}$ is such that $t_i \neq t_j$ for some i, j. Then there is a permutation $\mathbf{t}'$ of $\mathbf{t}$ such that $\mathbf{t}' \neq \mathbf{t}$. Because $w_L(\mathbf{t}) = N(\prod_{l=1}^{L-1} D(s,t_l))/s^{L-1}$ is invariant to permutations of its arguments, $w_L(\mathbf{t}') = w_L(\mathbf{t})$. Let $\mathbf{t}'' = (1/2)\mathbf{t} + (1/2)\mathbf{t}'$. By the strict convexity of $w_L(\mathbf{t})$, $w_L(\mathbf{t}'') < (1/2)w_L(\mathbf{t}) + (1/2)w_L(\mathbf{t}') = w_L(\mathbf{t})$. Furthermore, $\sum_{l=1}^{L-1} t''_l = \sum_{l=1}^{L-1} t'_l = \sum_{l=1}^{L-1} t_l$. Therefore, $DL(\mathbf{t}'') < DL(\mathbf{t})$, and $\mathbf{t}$ does not minimize DL.

To prove (b), note that $w_L(\mathbf{t})$ is strictly convex if $\log(w_L(\mathbf{t}))$ is strictly convex. The latter condition holds if $\log(D(s,t))$ is strictly convex in t, that is, if $D_t(s,t)/D(s,t)$ is strictly increasing in t. *Q.E.D.*

Proof of Example 1. In this case, for any given $L \geq 2$ and $\bar{s} \geq s \geq 2$, the problem of the hierarchy becomes

$$\min_{\mathbf{t}} \eta N \prod_{l=1}^{L-1} s^{1-t_l^\alpha} + \sum_{l=1}^{L-1} s^{l-1} t_l + \lambda \left(\eta N (1/s^{L-1}) \prod_{l=1}^{L-1} s^{1-t_l^\alpha} + \sum_{l=1}^{L-1} t_l \right). \tag{A2}$$

Assume interior solutions and let $\mathbf{t}^* = (t_1^*, \ldots, t_{L-1}^*)$ solve (A2). The first-order conditions with respect to t_k and t_{k+1} yield

$$s^{k-1} + \lambda = \eta N \left(1 + \frac{\lambda}{s^{L-1}}\right) \prod_{l=1}^{L-1} s^{1-(t_l^*)^\alpha} \ln(s)\alpha(t_k^*)^{\alpha-1}$$

$$s^{k} + \lambda = \eta N \left(1 + \frac{\lambda}{s^{L-1}}\right) \prod_{l=1}^{L-1} s^{1-(t_l^*)^\alpha} \ln(s)\alpha(t_{k+1}^*)^{\alpha-1}.$$

Rearranging these conditions, one obtains $\frac{t_k^*}{t_{k+1}^*} = (\frac{s^k+\lambda}{s^{k-1}+\lambda})^{\frac{1}{1-\alpha}}$. Because $\alpha < 1$ and $s \geq 2$, it follows that $t_k^* > t_{k+1}^*$. Furthermore, $\lim_{\lambda\to\infty} \frac{t_k^*}{t_{k+1}^*} = 1$ and $\partial(\frac{t_k^*}{t_{k+1}^*})/\partial\lambda < 0$. To check that the problem of the hierarchy is strictly convex and hence that $\mathbf{t}^*$ is the unique minimum, note that $D_t(s,t)/D(s,t) = -\alpha t^{\alpha-1} \ln(s)$ is strictly increasing in t. Part (b) of Proposition 3(ii) then implies that $\prod_{l=1}^{L-1} s^{1-t_l^\alpha}$ is strictly convex in t. The rest is routine. *Q.E.D.*

[27] I am very grateful to an anonymous referee whose extensive suggestions led to more concise and elegant proofs, especially those of Proposition 3(ii) and Example 1.

We now highlight some features of the model that are key to prove Propositions 3 and 4. Let

$$P(\mathbf{s},\mathbf{t}) \equiv \eta(w_L)N\prod_{l=1}^{L-1} D(s_l,t_l) \text{ and } DL(\mathbf{s},\mathbf{t}) \equiv \eta(w_L)w_L + \sum_{l=1}^{L-1} t_l$$

denote, respectively, total information-processing time and delay as a function of $\mathbf{s}$ and $\mathbf{t}$. Furthermore, let

$$V(\mathbf{s},\mathbf{t}) \equiv \sum_{l=1}^{L-1} n_l t_l = \sum_{l=1}^{L-1}\left(\prod_{k=0}^{l-1} s_k\right) t_l$$

denote total coordination time.

Claim A1. $P(\mathbf{s},\mathbf{t})$ and $DL(\mathbf{s},\mathbf{t})$ (and hence the cost of delay) are symmetric, that is, invariant to permutations of pairs (s_l, t_l) in $(\mathbf{s},\mathbf{t})$.

Proof. Obvious.

Claim A2. Suppose $(\mathbf{s},\mathbf{t})$ is such that, for some $i<j$, either $t_i < t_j$ or $s_i > s_j$, or both. Then total coordination time $V(\mathbf{s},\mathbf{t})$ can be strictly reduced by swapping t_i with t_j if $t_i < t_j$ or s_i with s_j if $s_i > s_j$.

Proof. The proof is divided into three parts.

(i) Suppose $(\mathbf{s},\mathbf{t})$ is such that, for some $i<j$, $t_i < t_j$. Let $(\mathbf{s}',\mathbf{t}')$ be such that $(s'_l,t'_l)=(s_l,t_l)$ for all $l\neq i,j$, $(s'_i,t'_i)=(s_i,t_j)$, and $(s'_j,t'_j)=(s_j,t_i)$. Then $V(\mathbf{s},\mathbf{t}) > V(\mathbf{s}',\mathbf{t}')$ because

$$\prod_{k=0}^{i-1} s_k t_i + \prod_{k=0}^{j-1} s_k t_j > \prod_{k=0}^{i-1} s_k t_j + \prod_{k=0}^{j-1} s_k t_i.$$

(ii) Suppose $(\mathbf{s},\mathbf{t})$ is such that, for some $i<j$, $s_i > s_j$. Let $(\mathbf{s}',\mathbf{t}')$ be such that $(s'_l,t'_l)=(s_l,t_l)$ for all $l\neq i,j$, $(s'_i,t'_i)=(s_j,t_i)$, and $(s'_j,t'_j)=(s_i,t_j)$. Then $V(\mathbf{s},\mathbf{t}) > V(\mathbf{s}',\mathbf{t}')$ because

$$s_i \sum_{l=i}^{j}\left(\prod_{k=0,k\neq i}^{l-1} s_k\right) t_l > s_j \sum_{l=i}^{j}\left(\prod_{k=0,k\neq i}^{l-1} s_k\right) t_l.$$

(iii) Suppose $(\mathbf{s},\mathbf{t})$ is such that, for some $i<j$, $t_i \le t_j$, and $s_i \ge s_j$, with at least one strict inequality. Let $(\mathbf{s}',\mathbf{t}')$ be a permutation of $(\mathbf{s},\mathbf{t})$ such that $(s'_l,t'_l)=(s_l,t_l)$ for all $l\neq i,j$, $(s'_i,t'_i)=(s_j,t_j)$, and $(s'_j,t'_j)=(s_i,t_i)$. Then (i) and (ii) imply that $V(\mathbf{s},\mathbf{t}) > V(\mathbf{s}',\mathbf{t}')$. *Q.E.D.*

Proof of Proposition 3. Let $(\mathbf{s},\mathbf{t})$ be such that, for some $i<j$, $t_i \le t_j$, and $s_i \ge s_j$, with at least one strict inequality. Consider a permutation of $(\mathbf{s},\mathbf{t})$ where (s_i,t_i) and (s_j,t_j) are swapped. By Claim A1, such permutation has no effect on total information-processing time and the cost of delay. By Claim A2(iii), such permutation strictly decreases total coordination time. Thus, $(\mathbf{s},\mathbf{t})$ is not optimal. *Q.E.D.*

Proof of Proposition 4. Let $(\mathbf{s},\mathbf{t})$ be a candidate solution to (6) and suppose $i<j$. Consider the following three cases:

(a) $t_i < t_j$ and $s_i > s_j$. By Proposition 3, $(\mathbf{s},\mathbf{t})$ cannot be optimal.
(b) $t_i < t_j$ and $s_i \le s_j$. Rearrange $(\mathbf{s},\mathbf{t})$ so that the only change is that t_i and t_j are swapped. By Claim A2(i), such rearrangement strictly decreases total coordination time. Total information-processing time and the cost of delay also (weakly) decrease if $D(s_i,t_i)D(s_j,t_j) \ge D(s_i,t_j)D(s_j,t_i)$ for all $t_i<t_j$ and $s_i \le s_j$, that is, if $D(s,t)$ is log-supermodular.
(c) $t_i \ge t_j$ and $s_i > s_j$. Rearrange (s,t) so that the only change is that s_i and s_j are swapped. By Claim A2(ii), such rearrangement strictly decreases total coordination time. Total information-processing time and the cost of delay also (weakly) decrease if $D(s_i,t_i)D(s_j,t_j) \ge D(s_i,t_j)D(s_j,t_i)$ for all $t_i \ge t_j$ and $s_i > s_j$, that is, if $D(s,t)$ is log-supermodular.

Thus, the log-supermodularity of $D(s,t)$ guarantees that in the optimum, $t_i \ge t_j$ and $s_i \le s_j$ for all $i<j$. *Q.E.D.*

Proof of Proposition 7. Formally, we have to show that if Assumption (S) holds and $L_W^+ \ge L_D^-$, then $\arg\min_{(s,z)\in S'} T(s,-z;\lambda)$ is increasing in the strong set order in λ, where $S'=\{(s,z):(s,-z)\in S\}$. The proof is divided into a number of steps.

Step 1. For fixed $\bar{s} \ge s \ge 2$, $\arg\min_{L\in S_L} T(s,L;\lambda) \in [L_D^-(s), L_W^+(s)]$.

Proof. Suppose by contradiction that there exists an $\hat{L} \in \arg\min_{L\in S_L} T(s,L;\lambda)$ such that $\hat{L} > L_W^+(s)$ (the case when $\hat{L} < L_D^-(s)$ is similar). Clearly $WB(s,\hat{L}) > WB(s,L_W^+(s))$ by definition of $L_W^+(s)$. $DL(s,\hat{L}) \ge DL(s,L_W^+(s))$ follows from the fact that DL is increasing in L for $L \ge L_D^-(s)$ (by quasi-convexity) and $\hat{L} > L_W^+(s) \ge L_D^-(s)$. Thus $WB(s,\hat{L}) + \lambda C(DL(s,\hat{L})) > WB(s,L_W^+(s)) + \lambda C(DL(s,L_W^+(s)))$. But this implies $\hat{L} \notin \arg\min_{L\in S_L} T(s,L;\lambda)$, a contradiction.

Step 2. Let $V=\{(s,L)\in\mathbb{N}\times\mathbb{N} : \bar{s}\ge s\ge 2,\ L\in[L_D^-(s),L_W^+(s)]\}$ and $V'=\{(s,z):(s,-z)\in V\}$. Then $\arg\min_{(s,z)\in S'} \times T(s,-z;\lambda) = \arg\min_{(s,z)\in V'} T(s,-z;\lambda) = \arg\max_{(s,z)\in V'} -T(s,-z;\lambda)$.

Proof. The first equality follows from Step 1, the second is obvious.

Step 3. V' is a lattice.

Proof. Note that $L_D^-(s)$ is decreasing in s because $\min_{L\in S_L} DL(s, L) = \max_{L\in S_L} -DL(s, L)$ and $-DL$ is submodular in (s, L) on S by Assumption (S). Similarly, $L_W^+(s)$ is decreasing in s. Thus, $-L_D^-(s)$ and $-L_W^+(s)$ are increasing in s, and therefore V' is a lattice.

Step 4. $-T(s, -z; \lambda)$ is supermodular in (s, z) on V' and satisfies increasing differences in $(s, z; \lambda)$ on V'.

Proof. $-WB(s, L)$ and $-DL(s, L)$ are supermodular in $(s, -L)$ on $S \supset V$ by Assumption (S). Because sums of supermodular functions are supermodular, it suffices to show that $-C(DL(s, L))$ is supermodular in $(s, -L)$ on V, where C is a strictly increasing and concave function. By supermodularity of DL in (s, L) on V,

$$DL(s + 1, L + 1) - DL(s + 1, L) \geq DL(s, L + 1) - DL(s, L).$$

Note that $DL(s, L)$ is decreasing in s and increasing in L on V because $L \geq L_D^-(s)$ and quasi-convexity in L. Let $\Delta = DL(s, L + 1) - DL(s, L) \geq 0$ and $\Delta + z = DL(s + 1, L + 1) - DL(s + 1, L), z \geq 0$. We have

$$\begin{aligned} C(DL(s + 1, L + 1)) - C(DL(s + 1, L)) &= C(DL(s + 1, L) + \Delta + z) - C[DL(s + 1, L)] \\ &\geq C(DL(s + 1, L) + \Delta) - C(DL(s + 1, L)) \\ &\geq C(DL(s, L) + \Delta) - C(DL(s, L)) \\ &= C(DL(s, L + 1)) - C(DL(s, L)), \end{aligned}$$

where the first inequality follows from the fact that C is strictly increasing and the second from the fact that DL is decreasing in s and the concavity of C. Thus, $C(DL(s, L))$ is supermodular in (s, L) on V and hence $-T(s, -z; \lambda)$ is supermodular in (s, z) on V'.

It remains to be checked that for all $(s, z) \in V'$ such that $(s'', z'') \geq (s', z')$, $-\{WB(s'', -z'') + \lambda C(DL(s'', -z''))\} + \{WB(s', -z') + \lambda C(DL(s', -z'))\}$ is increasing in λ. This is true provided $DL(s', -z') - DL(s'', -z'') \geq 0$, or equivalently

$$[DL(s', -z') - DL(s'', -z')] + [DL(s'', -z') - DL(s'', -z'')] \geq 0.$$

Note that $DL(s', -z') - DL(s'', -z') \geq 0$ because DL is decreasing in s. $DL(s'', -z') - DL(s'', -z'') \geq 0$ follows from the fact that $L_D^-(s) \leq -z'' \leq -z' \leq L_W^+(s)$ and DL is increasing in L on $[L_D^-(s), L_W^+(s)]$.

Step 5. Steps 2–4 ensure that all the conditions of Topkis's monotonicity theorem are fulfilled. Proposition 7 thus follows. *Q.E.D.*

Remark on Proposition 7. As mentioned in Section 5, the assumptions used to prove Proposition 7 can be modified to cover the case where $DL(s, L)$ is not always decreasing in s. In fact, one could assume that (i) for fixed $L \geq 2$, $WB(s, L)$ and $DL(s, L)$ are quasi-convex in s and (ii) $s_D^+ \geq s_W^-$, where s_D^+ and s_W^- are defined as follows. Fix $L \geq 2$ and let $S_s = \{s \in \mathbb{N} : \bar{s} \geq s \geq 2\}$. Let $s_W^-(L)$ be the smallest $\arg\min_{s\in S_s} WB(s, L)$ and $s_D^+(L)$ be the greatest $\arg\min_{s\in S_s} DL(s, L)$. Write $s_D^+ \geq s_W^-$ if $s_D^+(L) \geq s_W^-(L)$ for all $L \geq 2$. The proof of Proposition 7 can then be modified by first showing that, for fixed $L \geq 2$, $\arg\min_{L\in S_s} T(s, L; \lambda) \in [s_D^-(L), s_D^+(L)]$ and then defining a new set $V = \{(s, L) \in \mathbb{N} \times \mathbb{N} : s \in [s_W^-(L), s_D^+(L)], L \in [L_D^-(s), L_W^+(s)]\}$. Steps 2–4 can also be easily adapted.

Proof of Example 2. Fix s. Note that by quasi-convexity, $DL(s, L)$ is increasing in L for $L \geq L_D^-(s)$ and strictly decreasing in L for $L < L_D^-(s)$ and therefore $DL(s, L + 1) \geq DL(s, L) \Leftrightarrow L \geq L_D^-(s)$. $DL(s, L_W^+ + 1) \geq DL(s, L_W^+)$ can be written as

$$\bar{\tau} + b \geq N\left[\left(\frac{d(s)}{s}\right)^{L_W^+-1} \eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+-1}\right) - \left(\frac{d(s)}{s}\right)^{L_W^+} \eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+}\right)\right]. \tag{A3}$$

Thus (A3) implies $L_W^+ \geq L_D^-$. (For notational ease, we will often omit the reference to s in $L_W^+(s)$ and $L_D^-(s)$ in the following.)

Furthermore, by definition of L_W^+ it must be that $WB(s, L_W^+ + 1) > WB(s, L_W^+)$:

$$s^{L_W^+-1}\bar{\tau} > N\left[d(s)^{L_W^+-1}\eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+-1}\right) - d(s)^{L_W^+}\eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+}\right)\right]. \tag{A4}$$

Dividing the above inequality by $s^{L_W^+-1}$ yields

$$\bar{\tau} > N\left[\left(\frac{d(s)}{s}\right)^{L_W^+-1} \eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+-1}\right) - s\left(\frac{d(s)}{s}\right)^{L_W^+} \eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+}\right)\right]. \tag{A5}$$

Thus a sufficient condition for (A3) (and hence $L_W^+(s) \geq L_D^-(s)$) is that

$$b \geq (s - 1)N\left(\frac{d(s)}{s}\right)^{L_W^+} \eta\left(N\left(\frac{d(s)}{s}\right)^{L_W^+}\right). \tag{A6}$$

Now let $\eta(w) = \phi w^{1+\psi}$. (A5) can be rewritten as

$$\bar{\tau} > N^{(2+\psi)}\phi\left(\frac{d(s)}{s}\right)^{L_W^+(2+\psi)}\left[\left(\frac{s}{d(s)}\right)^{(2+\psi)} - s\right].$$

Furthermore, $d(s) = \beta s^{\alpha}$ implies $N(\frac{d(s)}{s})^{L_W^+} = N(\beta s^{\alpha-1})^{L_W^+}$. Thus, assuming that $(1-\alpha)(1+\psi) > \alpha$, we get

$$N(\beta s^{\alpha-1})^{L_W^+} < \left[\frac{\bar{\tau}}{\phi s\left[\left(\frac{1}{\beta}\right)^{(2+\psi)} s^{(1-\alpha)(1+\psi)-\alpha} - 1\right]}\right]^{\frac{1}{2+\psi}}. \tag{A7}$$

Manipulation of (A6) and (A7) yields the following sufficient condition for $L_W^+(s) \geq L_D^-(s)$:

$$b \geq \frac{(s-1)}{s}\frac{\bar{\tau}}{\left(\frac{1}{\beta}\right)^{(2+\psi)} s^{(1-\alpha)(1+\psi)-\alpha} - 1}$$

for all $s = 2, \ldots, \bar{s}$. Noting that $\frac{s-1}{s} \leq 1$ and that $s^{(1-\alpha)(1+\psi)-\alpha}$ is increasing in s yields a stronger condition:

$$b \geq \frac{\bar{\tau}}{\left(\frac{1}{\beta}\right)^{(2+\psi)} 2^{(1-\alpha)(1+\psi)-\alpha} - 1}.$$

Because $\beta < 1$, the claim in Example 2 follows. *Q.E.D.*

References

Acemoglu, D., Aghion, P., Lelarge, C., Van Reenen, J., and Zilibotti, F. "Technology, Information and the Decentralization of the Firm." *Quarterly Journal of Economics,* Vol. 122 (2007), pp. 1759–1799.

Becker, G.S. and Murphy, K.M. "The Division of Labor, Coordination Costs and Knowledge." *Quarterly Journal of Economics,* Vol. 107 (1992), pp. 1137–1160.

Bolton, P. and Dewatripont, M. "The Firm as a Communication Network." *Quarterly Journal of Economics,* Vol. 109 (1994), pp. 809–839.

Bureau of the Census. "First Findings from the EQW National Employer Survey." EQW Catalog no. RE01, 1995.

Calvo, C.A. and Wellisz, S. "Supervision, Loss of Control and the Optimal Size of the Firm." *Journal of Political Economy,* Vol. 86 (1978), pp. 943–952.

Chandler, A.D. *Strategy and Structure*. Cambridge, Mass.: The MIT Press, 1962.

——. *Scale and Scope: The Dynamics of Industrial Capitalism*. Cambridge, Mass.: Harvard University Press, 1990.

Galbraith, J.R. *Organization Design*. Reading, Mass.: Addison-Wesley, 1977.

Garicano, L. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy,* Vol. 108 (2000), pp. 874–904.

Geanakoplos, J. and Milgrom, P. "A Theory of Hierarchies Based on Limited Managerial Attention." *Journal of the Japanese and International Economies,* Vol. 5 (1991), pp. 205–225.

Guadalupe, M. and Wulf, J. "The Flattening Firm and Product Market Competition: The Effect of Trade Liberalization." Working Paper, Harvard Business School, 2008.

Guetzkow, H. "Communications in Organizations." In P.C. Nystrom and W.H. Starbuck, eds., *Handbook of Organizational Design*. Oxford: Oxford University Press, 1981.

Herbold, R.J. "Inside Microsoft: Balancing Creativity and Discipline." *Harvard Business Review,* Vol. 80 (2002), pp. 73–79.

Keren, M. and Levhari, D. "The Optimum Span of Control in a Pure Hierarchy." *Management Science,* Vol. 25 (1979), pp. 1162–1172.

—— and ——. "Decentralization, Aggregation, Control Loss and Costs in a Hierarchical Model of the Firm." *Journal of Economic Behavior and Organization,* Vol. 11 (1989), pp. 213–236.

Mahoney, T.A., Jerdee, T.H., and Carroll, S.J. "The Job(s) of Management." *Industrial Relations,* Vol. 4 (1965), pp. 97–110.

Meagher, K.J. "Generalizing Incentives and Loss of Control in an Optimal Hierarchy: The Role of Information Technology." *Economic Letters,* Vol. 78 (2003), pp. 273–280.

——, Orbay, H., and Van Zandt, T. "Hierarchy Size and Environmental Uncertainty." In M.R. Sertel and S. Koray, eds., *Advances in Economic Design*. Berlin Heidelberg: Springer-Verlag, 2003.

Mendelson, H. "Organizational Architecture and Success in the Information Technology Industry." *Management Science,* Vol. 46 (2000), pp. 513–529.

Orbay, H. "Information Processing Hierarchies." *Journal of Economic Theory,* Vol. 105 (2002), pp. 370–407.

PRAT, A. "Hierarchies of Processors with Endogenous Capacity." *Journal of Economic Theory,* Vol. 77 (1997), pp. 214–222.

QIAN, Y. "Incentives and Loss of Control in an Optimal Hierarchy." *Review of Economic Studies,* Vol. 61 (1994), pp. 527–544.

RADNER, R. "The Organization of Decentralized Information Processing." *Econometrica,* Vol. 61 (1993), pp. 1109–1146.

RAJAN, R.G. AND WULF, J. "The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies." *Review of Economics and Statistics,* Vol. 88 (2006), pp. 759–773.

SAYLES, L.R. *Managerial Behavior*. New York: McGraw-Hill, 1964.

STARBUCK, W.H. "Organization Growth and Development." In W.H. Starbuck, ed., *Organizational Growth and Development*. Harmondsworth, UK: Penguin Books, 1971.

TING, M.M. "A Strategic Theory of Bureaucratic Redundancy." *American Journal of Political Science,* Vol. 47 (2003), pp. 274–292.

TOPKIS, D.M. *Supermodularity and Complementarity*. Princeton, NJ: Princeton University Press, 1998.

VAN ZANDT, T. "Continuous Approximations in the Study of Hierarchies." *RAND Journal of Economics,* Vol. 26 (1995), pp. 575–590.

——. "The Scheduling and Organization of Periodic Associative Computation: Efficient Networks." *Review of Economic Design,* Vol. 3 (1998a), pp. 93–127.

——. "Organizations with an Endogenous Number of Information Processing Agents." In M. Majumdar, ed., *Organizations with Incomplete Information*. Cambridge, UK: Cambridge University Press, 1998b.

——. "Real-Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents." *Review of Economic Studies,* Vol. 66 (1999), pp. 633–658.

——. "Structure and Returns to Scale of Real-Time Hierarchical Resource Allocation." Working Paper, INSEAD, 2003.

VAYANOS, D. "The Decentralization of Information Processing in the Presence of Interactions." *Review of Economic Studies,* Vol. 70 (2003), pp. 667–695.

WILLIAMSON, O.E. "Hierarchical Control and Optimum Firm Size." *Journal of Political Economy,* Vol. 75 (1967), pp. 123–138.