



**DEPARTMENT OF ECONOMICS**

**DISCUSSION PAPER SERIES**

**OPTIMAL COORDINATION IN HIERARCHIES**

**Andrea Patacconi**

Number 238

June 2005

Manor Road Building, Oxford OX1 3UQ

# Optimal Coordination in Hierarchies

Andrea Pataconi  
University of Oxford

June 2005

## Abstract

This paper studies the optimal allocation of coordination responsibilities in organizations where duplication of effort is a serious concern. The planner's objective is to minimize a weighted average of the wage bill and the cost of delay. The paper provides conditions under which, in balanced hierarchies, communication effort is increasing and the span of control is decreasing as one travels up the hierarchy, with equalities holding if wages are negligible relative to the weight attached to the cost of delay. The analysis suggests that concerns for fast decision-making may be key in explaining the recent trend towards empowerment in firms. Several variants of the basic model are studied, including one focusing on communicative skills and another in which, as urgency increases, the optimal span of control increases and the hierarchy flattens. Evidence supporting these results is discussed.

**Keywords:** Coordination, Hierarchy, Duplication, Delay, Information Processing.

**JEL Classification:** D21, L23

---

<sup>1</sup>Postal address: Andrea Pataconi, Department of Economics, Manor Road Building, Manor Road, Oxford, OX1 3UQ, United Kingdom. E-mail: andrea.pataconi@economics.ox.ac.uk.

I am particularly indebted to Meg Meyer for her guidance at every stage of the project and to John Quah for comments that have greatly improved the paper. I also would like to thank Dan Anderberg, Ernesto Dal Bo, Wouter Dessein, Erik Eyster, John Roberts, seminar participants at the University of Oxford, Royal Holloway and the 2004 Congress of the European Economic Association at the Universitat Carlos III in Madrid for helpful conversations and remarks. Financial support from the Economics Department at the University of Oxford and the Royal Economic Society is gratefully acknowledged. All remaining errors are mine.

# 1 Introduction

The division of labor, upon which the wealth of nations depends, requires the existence of means by which specialized activities can be properly coordinated. Within the firm, for instance, decisions are usually taken through a complex system of authority levels which is to a large extent a matter of planning. Indeed, one of the fundamental tasks of the entrepreneur (or the management in large organizations) is to ensure that the activities within the firm are properly organized and coordinated.

This paper is concerned with the optimal organization of coordination activities within the firm – identified here as a hierarchy. Besanko et al. (2000) argue that "coordination involves the flow of information to facilitate subunit decisions that are consistent with each other and with organizational objectives" (p.549). This paper deals with a specific but important aspect of that definition, namely the efficient division of labor within hierarchical organizations. A key assumption I put forward is that by spending more effort on coordination and planning activities such as work scheduling, arranging meetings or the exchanging of information with colleagues, managers can reduce overlaps among tasks and therefore wasteful duplications of effort.<sup>1</sup>

Duplication of effort has long been recognized as a serious concern, especially in large organizations and the public sector (Simon et al. (1950)). Chandler (1990), for instance, reports that, by the end of 1920, the director of Du Pont's Chemical Department, together with the departmental research directors, took on the responsibility of coordinating the different manufacturing departments "so that overlapping of the research programs [...] may be avoided as completely as possible" (p.182). Similarly, Stuart Rice, in his analysis of the role and management of the Federal Statistical System, rhetorically asks why "if I head an electric utility, should I be asked to report my total pay-roll or number of employees, or both, to six different federal offices, on 12 different forms, at intervals varying from one month to one year?" (1940, p.482). He emphasizes, as this paper does, that duplication of effort usually stems from the decentralization of agencies (i.e., the division of labor) and that coherence in organization can be attained only by an item to item adjustment of each task and process to every other related task and process, that is, by coordinating interrelated activities.

The recent wave of corporate investments in information technology is also in part motivated by the need to reduce duplications of effort. In the pharmaceutical industry, for instance, companies test huge numbers of different compounds in search for new drugs, and information systems such as Intranets are routinely used to share information, improve collaboration among organization units and minimize duplications of research effort. Two examples among many are SmithKline Beecham, which runs virtual

---

<sup>1</sup>Of course in reality not all duplications need to be wasteful. Many authors, for instance, have argued that redundancies may actually be an optimal response to the possibility of 'misunderstandings' in the communication process (e.g., Chwe (1995)). However, in my final remarks, I will argue that the present model can be reinterpreted along these lines.

libraries that can be consulted by each of its employees, and Glaxo-Wellcome, whose research scientists can access complex chemical information including structural images on their internal Intranet (see [www.skyrme.com/insights/25intra.htm](http://www.skyrme.com/insights/25intra.htm) for more examples from different industries).

Notwithstanding its practical relevance, it is fair to say that the problem of duplication of effort has received little attention in the economics literature.<sup>2</sup> Previous work, in fact, has focused on other aspects of ‘loss of control’ in organizations, most notably the loss of useful information (e.g., Williamson (1967) and Vayanos (2003)) and shirking (e.g., Calvo and Wellisz (1978,1979) and Qian (1994)). This paper extends this literature by exploring the implications of a tradeoff between duplication of effort and coordination costs in hierarchical organizations that must carry out a task of given size. The organization’s task can be interpreted either as producing goods such as cars or drugs or as processing information (see below). In line with previous work, upper-level managers can delegate their (sub)tasks to a few subordinates, whom I call a workgroup, in order to reduce delay. However, in this model, superiors must also spend time and effort coordinating their subordinates if they want to reduce wasteful duplications of tasks and effort.<sup>3</sup> More specifically, I assume that the percentage of a delegated task which is duplicated, captured by the function  $D(n, c)$ , depends on both  $n$ , the number of subordinates a manager has (that is, his span of control) and the superior’s communication effort  $c$  (also called coordination effort). Thus,  $1/D(n, c)$  can be interpreted as a measure of coordination at a given layer. Importantly, since both communication effort and span of control will be decision variables at each level of the hierarchy, I will be able to compare their optimal levels across layers.

The focus will be on two scenarios, which are termed "production hierarchies" and "information processing hierarchies". In the basic model, a neat distinction is drawn between bottom-level employees (or workers), who are engaged in production, and upper-level managers, whose only job is to coordinate their direct subordinates. In this scenario the task is best interpreted as the production of goods, so the organization is termed a production hierarchy. Many well-known models share this basic structure, including Calvo and Wellisz (1979), Keren and Levhari (1979) and Qian (1994).

In information processing hierarchies, by contrast, performing the task requires the top manager to acquire some relevant information. As in the production scenario, upper-level managers are in charge of coordinating the work of their subordinates and raw information is processed only by bottom-level managers. However, reports summarizing the relevant information must now be transmitted up the hierarchy to the top management, as in Radner (1993). Thus, upper-level managers will spend time both processing information (i.e., reading reports) and coordinating. This scenario is studied in Section 6.

I stress that the crucial distinction between production and information processing hierarchies revolves

---

<sup>2</sup> A notable exception is Bolton and Farrell (1990); their focus however is not on the internal organization of firms.

<sup>3</sup> Alternatively, subordinates may coordinate their work autonomously (“horizontal coordination”). As explained in Section 2, the model could be easily adapted to cover this case.

around whether or not upward reporting takes place within the organization. In particular, it may well be possible that the task involves some processing of information even in production hierarchies, so long as this information is not communicated to the senior management. Thus, for consistency and expositional ease, I will usually refer to the task as the processing of information, both in the production and the information processing scenario.

A second important remark is that, in both scenarios, attention will be restricted to hierarchies in which all employees at the same level are treated similarly. More precisely, I will require that tasks are divided evenly among the members of the same workgroup. This in fact is sufficient, in my setup, to guarantee that hierarchies are balanced, with managers at the same level choosing the same coordination effort and span of control.<sup>4</sup>

This paper studies how coordination responsibilities should be optimally allocated across layers taking into account the simple but important fact that communicating is inherently costly. The time managers spend on coordinating (which I term coordination costs), in fact, has to be computed as part of the organization's total working time and taken into account at the moment of making decisions. The organization's objective is to minimize a weighted average of the total wage bill and the cost of delay. The optimal level of communication results from a basic tradeoff between the savings on processing time due to less duplication and coordination costs. I stress that, as in the team-theoretic literature, no conflict of interest is assumed between the organization and its members, and managers will therefore simply maximize the organization's objective.

Although my analysis neglects the crucial role of incentives in organizations, it emphasizes a number of other important issues in the design of organizations, most notably the role of delay. To build some intuition, I first consider a simplified framework in which the span of control is constant across layers (i.e., a uniform hierarchy). In that setup, I show that in the optimum higher-level managers spend more time on communicating and coordinating than lower-level managers do. In production hierarchies, this is simply due to the fact that the effort of a senior manager influences a much greater portion of the hierarchy than the effort of a lower-level manager. Coordination is therefore more cheaply provided by higher-level managers. However, when managers must also transmit their information to their superiors (as in Section 6), a second asymmetry arises. In fact, provided that duplications made by subordinates can be detected and disregarded at no cost by their superiors, a subordinate's coordination efforts do not reduce (in form of shorter reports) his superior's information processing workload. By contrast, duplications made at higher levels always trickle down the hierarchy through the delegation process. Again, coordination effort is most valuable when exerted by the senior management, but now the reason is also that the benefits of coordination are higher at the top of the organization.

Empirically, the prediction that the time a manager spends on coordination activities increases as one

---

<sup>4</sup>The role of this assumption is discussed further in Section 2, where the model is presented.

moves up the hierarchy is strongly supported by the evidence. Mahoney et al. (1965), for instance, find that the percentage of managers who plan and the percentage of managers who coordinate as their main function increase as one passes from low to middle levels in the hierarchy, and then again from middle to high.<sup>5</sup> Similarly, several other studies (see the surveys by Sayles (1964) and Guetzkow (1981), and the references therein) support the view that the managerial job entails a continuing effort to coordinate activities within the organization, and that planning and coordinating receive the greatest emphasis within top management. Importantly, these findings cannot entirely be attributed to the well-known fact that managers work longer hours the higher they are in the hierarchy, since they typically show that it is the percentage of time spent on planning and coordinating to total hours worked that is increasing with rank. Rather, the evidence suggests that the longer hours spent on these activities might be one of the reasons why higher-level managers work longer hours compared to lower-level managers.

A second result of this paper is that a shift towards granting employees broader decision authority – on the way work is coordinated, in the present model – has to be expected when reducing delay becomes more important. More precisely, the model shows that the ratio of communication effort exerted by a subordinate to the coordination effort exerted by his superior tends to increase with urgency and in the limit (i.e., when only delay matters) communication efforts are equalized across (managerial) layers. In this precise sense, therefore, increased urgency tends to ‘empower’ lower-level managers, relative to their superiors. Interestingly, the result seems broadly consistent with the view that innovative firms have more decentralized hierarchical structures than less innovative companies, and that globalization, by increasing competition (and therefore the need for fast decision-making), has had a major impact in forcing firms to restructure (Marin and Verdier (2003)).

Section 4 studies the more general case of balanced hierarchies, where the choice of the span of control across layers is also endogenous. Smaller spans of control near the top of the organization are shown to be typically beneficial as they reduce the number of managers in the hierarchy (keeping the number of workers fixed) and therefore coordination costs, particularly near the top where communication requirements are higher. Sufficient conditions are provided for the optimal span of control to be decreasing and coordination effort increasing as one travels up the hierarchy, with equalities holding if wages are negligible relative to the marginal cost of delay, thus generalizing the main result of Keren and Levhari (1979). These conditions turn out to be quite intuitive and, loosely speaking, essentially require communication effort and the span of control not to be ‘too complementary’ at reducing duplications. Furthermore, as for the time devoted

---

<sup>5</sup>Managerial jobs which are usually referred to as coordination activities by the empirical literature on human resource management (HRM) include the exchanging of information with people in the organization in order to relate and adjust programs, advising other departments, arranging meetings, etc. Planning activities such as work scheduling and programming are also examples of “coordination” as defined in the present model. The HRM literature also stresses the importance (in terms of time) of the planning and coordinating functions in actual organizations (e.g., Mahoney et al. (1965)).

to coordination activities, there seems to be some empirical support for the fact that the span of control is smaller near the top of the organization (e.g., Starbuck (1971), Gabraith (1977)), and Keren and Levhari (1979) also argue that, in military organizations, where wage costs are plausibly secondary compared to the utility of planning time saved, spans of control tend to be more uniform.

I also wish to emphasize the methodological contribution that this paper makes to the literature. Previous work on loss of control, in fact, including Keren and Levhari (1979,1983,1989), Qian (1994) and Meagher (2003), has relied heavily on the use of continuous approximations in the study of hierarchies. Van Zandt (1995), however, has shown that these approximations can be inaccurate, especially so far as the length of the hierarchy is concerned. It is therefore worthy of mention that this paper extends previous results without relying on approximations but using instead interchange arguments which are standard in dynamic programming (see, e.g., Ross (1983)).

The present model can also be easily adapted to study the optimal choice of the managers' communicative skills in hierarchies. I consider a variant of the basic model in which managers devote the same amount of time to coordination activities, but the effectiveness of their effort depends on their ability. Unsurprisingly, managers with strong interpersonal and communicative skills will typically be located at the top echelons of the hierarchy. However, the model also predicts that, as urgency increases, the optimal skill level of all managers should increase, especially at the lower and middle levels of the hierarchy. Empirically, this result suggests that firms operating in turbulent environments (and for which delay is presumably very costly) should hire more skilled managers and provide more training opportunities to their employees, especially at the bottom of the hierarchy, than companies operating in traditional sectors.

Finally, I consider another variant of the basic model (with constant span of control across layers) incorporating both a concern for delay and gains from specialization in information processing. I first show that the span of control and the number of levels in the hierarchy are, in a specific technical sense, substitutes, as they both increase the ability of the organization to process information concurrently by expanding its size. The focus of the analysis is on how these two variables vary together in the optimum as urgency increases. I show that if, in a sense made precise in the paper, delegation is mainly driven by specialization, then the span of control increases and the hierarchy flattens as reducing delay becomes more important. In a specific example, this condition is more likely to be fulfilled when delegation involves large fixed delay costs, gains from specialization are substantial and coordination costs are small. Taken together, my results offer a first formalization of the widespread view among practitioners that empowerment and layering are to a large extent driven by the need for fast decision-making and, I hope, may shed some light on recent evidence on the changing nature of corporate hierarchies as documented, for instance, by Rajan and Wulf (2004).

## 1.1 Related Literature

This paper is related to several strands of the literature on hierarchies and organizations. It builds on recent contributions to the theory of information processing networks (e.g., Radner (1993), Bolton and Dewatripont (1994)), but the focus of the analysis is different. The information processing literature in fact focuses on the optimal design of the network and assumes that information arrives to the organization already disaggregated in the form of distinct batches of information. By contrast, the emphasis here is on coordinating the division of labor among workers and, for tractability, attention is restricted to a subset (albeit a very important one) of all the possible network configurations, namely the class of balanced hierarchies.

This paper most directly contributes to a large literature on loss of control in organization (e.g., Williamson (1967), Calvo and Wellisz (1978), Keren and Levhari (1979)). Relative to previous contributions, besides the original focus on duplications, the present work pays special attention to the issue of delay. Considerations of delay are undoubtedly key in shaping organizational decisions and indeed several recent trends in the design of organizations have been attributed to a need for faster decision making and execution. Yet, it is fair to say that, with a few important exceptions,<sup>6</sup> little work has been done to understand the impact of increased urgency on organizational design. This paper therefore fills an important gap in the literature and, in so doing, provides new insights on the issues of empowerment and layering.

Lastly, this paper is related to a recent and fast-growing literature on the role of coordination within the firm (e.g., Dessein and Santos (2003), Harris and Raviv (2002), Rotemberg (1999)). Within this literature, the contribution most closely related to the present one is Hart and Moore (1999), which shows that coordinators (i.e., individuals whose tasks cover a large subset of assets) should appear higher in the chain of command than specialists (i.e., those with a narrow remit). However, Hart and Moore analyze hierarchy in terms of authority (which means that if  $i$  is above  $j$  in the hierarchy, then necessarily  $i$  has authority over  $j$ ), whereas my work takes an information processing approach. The two papers are therefore best seen as complements rather than substitutes.

The remainder of the paper is organized as follows. Section 2 introduces the basic notation and assumptions of the model. Section 3 focuses on uniform hierarchies. Section 4 endogenizes the choice of the span of control and proves the main results of the paper. Section 5 focuses on layering, while Section 6 studies information processing hierarchies. Section 7 concludes. All the omitted proofs are gathered in the Appendix.

---

<sup>6</sup>Keren and Levhari (1979) is a case in point. Beggs (2001) also characterizes the optimal design of the hierarchy where a trade-off exists between cost and delay and highlights a subtle relationship between increased costs of delay and decentralization. He shows that increased urgency need not always result in more decentralization because reducing the number of levels tasks have to pass through is only one of the methods an organization can employ to reduce delay. Another method, which may be cheaper, is to hire more low-ability workers to reduce the time spent queueing for attention.



## 2 The Model

This section introduces the main innovation of the paper, namely the duplication function, and discusses the key assumptions of the paper. A simple two-layer hierarchy is studied to illustrate the basic tradeoffs of the model, and the general setup is then presented.

### 2.1 Duplication of Effort

Duplication of effort is modeled as follows. Consider a manager who must carry out a task of size  $N \in \mathbb{R}_{++}$ . As in Becker and Murphy (1992), it is assumed that any task can be subdivided into infinitely many subtasks. Suppose that the manager delegates and subdivide his task evenly among  $n$  subordinates (called a workgroup). Then duplications can occur as a result of a lack of coordination during the division of labor. Specifically, I posit that the actual total amount of information processed by the workers is not  $N$  but  $ND(n, c) \geq N$ , where  $n \in \mathbb{N}$  is the span of control of the manager and  $c \in [0, 1]$  denotes the total communication effort exerted by the manager. Thus,  $D(n, c)$  measures the amount of duplications in the organization or, equivalently, the lack of coordination among subordinates. Note also that the manager cannot retain part of the task for himself. This assumption is standard in models of loss of control and could be defended on the ground that managers have different skills than production workers,<sup>7</sup> but of course it would be worthwhile to relax it in future research.

Formally, the duplication function is defined as follows:

**Definition (DF)** For all  $c \in [0, 1]$  and  $n$ , the duplication function  $D : [0, 1] \times \mathbb{N} \rightarrow \mathbb{R}$  satisfies the following assumptions: (i)  $D(n, c)$  is twice continuously differentiable in  $c$ , (ii)  $D(n, c) \geq 1$ , (iii)  $-\infty < D_c(n, c) < 0$  for all  $n \leq \bar{n}$  and (iv)  $D(n, c) = n$  for all  $n > \bar{n}$ .<sup>8</sup>

The first three assumptions impose mild restrictions on  $D(\cdot, \cdot)$ . In particular, (ii) just states that we are dealing with duplications, while (iii) simply says that by communicating more managers can reduce duplication of effort. The fourth assumption is convenient in order to ensure that in the optimum the span of control is finite even in the extreme case where the organization only cares about delay, and could be easily relaxed. In particular, it would be enough to require duplications to be sufficiently large as group size grows beyond a certain level, an assumption which is broadly consistent with a large managerial literature postulating an upper bound to the number of subordinates that can be effectively supervised.<sup>9</sup> It should also be noted that this paper takes a reduced-form approach in modeling duplications. In the conclusion, however, I will sketch two possible ways in which the duplication function can be micro-founded.

<sup>7</sup>Throughout this paper, however, all agents are assumed to be identical.

<sup>8</sup>By convention, subscripts in functions of more than one variable will be used to denote partial derivatives.

<sup>9</sup>Urwick (1956), for instance, relates the optimality of limited spans of control to bounds to "the number of items that the human brain can keep within its grasp simultaneously" (p.41). Psychological research also shows that disruption of communication is more likely to occur as the size of the group increases (see, e.g., Guetzkow (1981)).

The simple framework outlined above can be extended to deal with multi-layer organizations. To fix ideas, consider a hierarchy composed of three layers, 1, 2 and 3, 1 being the top level and 3 the bottom level (the general framework is presented later). Let  $c_1, c_2$  and  $n_1, n_2$  be, respectively, the communication efforts and spans of control at layer 1 and 2. In line with the discussion above, let  $ND(n_1, c_1)$  be the size of the task the middle management believe they are in charge of (that is, the actual task plus the eventual duplications). Clearly, if tasks are divided evenly, each of the middle-level managers will delegate a task of size  $ND(n_1, c_1)/n_1$  to his subordinates and, provided he exerts effort  $c_2$ , each worker's actual workload will be  $ND(n_1, c_1)D(n_2, c_2)/n_1n_2$ . Thus, workers' individual workloads exhibit a simple recursive structure which depends on their superiors' efforts only through the multiplicative term  $D(n_1, c_1)D(n_2, c_2)$ .

Some features of the model are worthy of further discussion. The most important remark concerns the assumption that tasks are divided evenly among subordinates. That assumption greatly simplifies the model and allows me to focus on balanced hierarchies. In fact, if all managers at the same layer face the task, one can assume with (essentially) no loss of generality that they will make the same decisions and, in particular, that they will choose the same levels of communication effort and the same span of control. The focus on balanced hierarchies is typically motivated by the fact that these organizational forms resemble, at least to a first approximation, actual managerial charts. Furthermore, in the present setting, their analytical simplicity allows me to focus on the main theme of the paper, namely coordination in organizations. Needless to say, it would nevertheless be desirable to fully endogenize the division of labour within the organization by assigning to each agent a task of a potentially different measure.<sup>10</sup>

A second feature of the model which must also be emphasized is that the duplication function is the same at every level of the hierarchy. This is consistent with the assumption that the organization's task is perfectly divisible and homogeneous. However, in reality the job of the senior management is often qualitatively different from the job of lower-level managers. In my final remarks, I will briefly mention how asymmetries in the difficulty/complexity of the task across layers may affect the conclusions of the paper.

Finally, I stress that although a superior only coordinates the division of labor within the workgroup he supervises, coordination among workgroups is guaranteed by the efforts of the upper-level managers. Thus, for instance, even if there was no duplication within workgroups at the bottom level of the hierarchy, there may still be duplications among workgroups as a result of poor coordination at previous layers.

## 2.2 A Two-Layer Hierarchy

In this subsection, I focus on the simple case of a two-layer hierarchy and use this setup to illustrate the basic tradeoffs of the model.

---

<sup>10</sup>In particular, Radner (1993) has shown, in a different setup, that the unbalanced hierarchies he termed 'reduced trees' are optimal. Interestingly, however, both that paper and Van Zandt (2004) provide conditions under which optimal hierarchies are 'approximately balanced'.

I start with the trivial observation that processing information takes time. Formally, let  $\eta w$  be the time spent by a manager processing an amount  $w$  of information (his workload), where  $\eta > 0$  clearly denotes the time necessary to process one unit of information. Thus, for instance, if a worker were to process all the information by himself (a one-layer hierarchy), he would spend  $\eta N$  units of time. The assumption that processing time is linear in  $w$  is only for simplicity: any increasing function of  $w$ , perhaps capturing gains from specialization or the cost of overloads, would leave all the qualitative results of the paper unchanged.<sup>11</sup>

Planning and coordinating the division of labor also take time. Specifically, let  $\tau(c)$  be the time that a manager must spend on coordinating to achieve a communication level  $c$ .<sup>12</sup>  $\tau(c)$  is assumed to be twice continuously differentiable, with  $\tau(0) = 0$  and  $\tau'(\cdot) > 0$ . The assumption that  $\tau(\cdot)$  is increasing simply reflects the fact that, if coordination has to be improved, the time spent on communicating must increase.

There are at least two ways in which  $\tau(c)$  can be interpreted (Subsection 4.1 provides a third interpretation in terms of ability). For instance, if  $c$  denotes *downward communication*, then  $\tau(c)$  measures the amount of time that a superior spends communicating with his subordinates. Alternatively,  $\tau(c)$  may be interpreted as the time each subordinate spends communicating with the other members of his workgroup, in which case  $c$  should be thought of as *lateral communication*. In the following, I will adopt the former interpretation; the results however would remain the same (up to some changes in notation) if we adopted the second interpretation.

In addition to the variable communication cost  $\tau(c)$ , delegation also involves a fixed time cost. In particular, an amount of time  $b > 0$  is assumed to elapse before a manager can communicate with his subordinates.  $b$  may for instance represent the time passing between when an e-mail is sent by a superior and when it is read by his subordinates, or the time necessary to organize and inform all participants about a meeting, whereas  $\tau(c)$  would measure, respectively, the time the superior spends writing that e-mail or attending that meeting.

The organization wishes to minimize the total cost of processing all the information. Two sources of costs are considered, the wage bill and the cost of delay. The wage bill includes the wages paid to the employees for the time they actually spend processing information and coordinating. In particular, as in Bolton and Dewatripont (1994), employees are not compensated for the time in which they remain idle, thus implicitly assuming that agents or the organization can use this idle time for other purposes. Delay is defined in the usual way as the time the organization takes to process all the information, taking into account that some tasks might be processed in parallel. Thus, for instance, in a one-layer hierarchy, the

---

<sup>11</sup>Most of the results would carry through even if  $\eta(w)$  is decreasing in  $w$ . In Section 5, however, additional assumptions would have to be imposed on  $\eta(w)$ .

<sup>12</sup>The time and effort it takes to the 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.

wage bill is  $\omega\eta N$  (where  $\omega$  denotes the wage for unit of time and with no loss of generality is normalized to unity in the sequel) and delay is given by  $\eta N$ . The total cost of processing an amount of information  $N$  using only one worker is given by  $\eta N + \lambda\eta N$ , where  $\lambda > 0$  denotes the weight attached to (the cost of) delay relative to wage costs (or, alternatively, the constant marginal cost of delay). I emphasize that the assumption that the cost of delay is linear in delay is just a simplification as any strictly increasing function would leave all the forthcoming results unchanged.

Now consider a two-layer hierarchy. In this scenario, the top manager (at layer 1) delegates the task to a number of subordinates (say  $n$ ) at layer 2. As we have seen, the size of the task at the second level will be  $ND(n, c)$  and  $\tau(c)$  is the time the top manager spends on coordinating. Since tasks are divided evenly among subordinates, a subordinate's workload is given by  $w = ND(n, c)/n$  and  $\eta ND(n, c)/n$  is the time he spends processing information. The problem of the two-layer hierarchy is

$$\min_{n,c} T(n, c) = [\eta ND(n, c) + \tau(c)] + \lambda [\eta ND(n, c)/n + \tau(c) + b]. \quad (1)$$

Three features of this problem are worth mentioning. First, note that delay includes only the time spent processing information by *one* manager at layer 2, since the managers at the same level in the hierarchy work concurrently (this is a standard assumption in the literature). Second,  $b$ , the time elapsing before a manager can communicate with his subordinates, only affects delay. This is a consequence of the fact that the organization does not pay its employees for the time they remain idle. Third, an equal division of tasks among subordinates is clearly optimal in this case, since delay is minimized. Should individual workloads differ, instead, delay would be given by time it takes the *slowest* worker to process his task (plus the communication costs  $\tau(c) + b$ ), which is clearly suboptimal.

As an aside, let me briefly consider how the problem of the hierarchy would change in an information processing scenario. Remember that, in that case, bottom-level employees would also have to transmit the information they have processed up to the top manager. Assume that a subordinate's report is proportional to the information he has processed. Thus, in our example,  $rND(n, c)/n$  would measure the size of the report that each subordinate sends to the top manager ( $r < 1$  is a parameter measuring information compression). Two polar cases can be considered. In the no-detection scenario, the top manager's workload is simply  $rND(n, c)$ : he just processes all the information he receives. In the detection scenario, by contrast, the top manager is assumed to be able to detect and disregard duplicated information at no cost; his workload is therefore only  $rN$ . Both these scenarios will be studied in Section 6.

Returning to our production hierarchy and assuming interior solutions, the first order condition (FOC) of (1) with respect to  $c$  yields<sup>13</sup>

$$(1 + \lambda)\tau'(c^*) = -\eta N(1 + \lambda/n)D_c(n, c^*) \quad (2)$$

---

<sup>13</sup>Throughout the paper, stars will denote optimal values.

which highlights a basic tradeoff between coordination costs and duplication of effort.

The span of control  $n$  and the length of the hierarchy  $L$  are also endogenous variables. Assuming just for the moment that the span of control can be treated as a continuous variable, from the first order condition with respect to  $n$  we get

$$-\lambda \left. \frac{\partial(D(n, c)/n)}{\partial n} \right|_{c=c^*, n=n^*} = D_n(n^*, c^*). \quad (3)$$

Thus, if  $\frac{\partial(D(n, c)/n)}{\partial n} < 0$ , the benefit of a larger span of control is that bottom-level managers have smaller workloads and therefore delay is reduced. The drawback is that duplications and hence wage costs may increase ( $D_n(n, c) > 0$ ). The latter effect can be seen as a formalization of Simon and al.'s insight that "the maximum size of an effective unit is limited basically by the ability of that unit to solve its problems of internal communication" (1950, p.131). Indeed,  $D_n(n, c) > 0$  seems a particularly natural assumption since, for given  $c$ , duplications will tend to rise with group size as the top manager spends on average less time communicating with each subordinate. (This effect may however be mitigated if several people interact simultaneously, as in a group meeting.) Furthermore, even when the amount of communication *per capita*  $\frac{c}{n}$  is fixed, each unit of communication may nevertheless become less effective as group size grows because misunderstandings and coordination failures in general are more likely in large groups (Guetzkow (1981)).<sup>14</sup>

Finally, the (optimal) two-layer hierarchy is preferred to a one-layer/one-manager hierarchy whenever the total cost of the former (wage bill and delay) is lower than the total cost of the latter:

$$\lambda N \eta \left[ 1 - \frac{D(n^*, c^*)}{n^*} \right] > N \eta [D(n^*, c^*) - 1] + (1 + \lambda) \tau(c^*) + \lambda b. \quad (4)$$

That is, it must be the case that the reduction in delay due to parallel information processing, which is proportional to  $1 - \frac{D(n^*, c^*)}{n^*}$ , outweighs duplication and coordination costs. Equation (4) is more likely to be satisfied when duplications and coordination costs are small and  $\lambda, \eta$  and  $N$  are large.

## 2.3 The Problem of the Hierarchy

The two-layer model studied above can be generalized to hierarchies with an arbitrary number of layers. Let the organization be composed of  $L \geq 2$  layers. Tasks are delegated as follows. The top manager, at level 1, delegates his tasks (of size  $N$ ) to  $n_1$  subordinates and exert communication effort  $c_1$ . Layer 2 total task (that is, the information managers at layer 2 believe they are responsible for) is  $ND(n_1, c_1)$ . Each manager at layer 2 may then delegate his tasks (the size of which is  $ND(n_1, c_1)/n_1$ ) to  $n_2$  managers at layer 3. Thus, the task of each workgroup at layer 3 is  $ND(n_1, c_1)D(n_2, c_2)/n_1$  ( $ND_1D_2/n_1$  for short). Layer 3 total and individual tasks are, respectively,  $ND_1D_2$  and  $\frac{ND_1D_2}{n_1n_2}$ . In general, layer  $l$  total task is  $N \prod_{i=1}^{l-1} D(n_i, c_i)$  and

---

<sup>14</sup>Let  $\tilde{D}(n, s) \equiv D(n, c)$ ,  $s = \frac{c}{n}$ . Then these two effects can be formally described by requiring  $\tilde{D}_s(n, s) < 0$ ,  $\tilde{D}_n(n, s) > 0$ .

layer  $l$  individual tasks is  $N \prod_{i=1}^{l-1} \frac{D(n_i, c_i)}{n_i}$ ,  $l = 2, \dots, L$ . At layer  $L$ , tasks cannot be delegated further and the information is processed. Thus, a worker's workload is given by  $w_L = N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ .<sup>15</sup>

Note that, since  $n_l$  is the span of control of the manager at layer  $l = 1, \dots, L-1$  (or, equivalently, the size of a workgroup at layer  $l+1$ ), the number of managers at layer  $l$  is  $\prod_{i=0}^{l-1} n_i$ . ( $n_0 = 1$  refers to the top manager.) Total processing time is therefore given by  $\eta \prod_{i=0}^{L-1} n_i w_L = \eta N \prod_{i=1}^{L-1} D(n_i, c_i)$ . Total working time (processing time and coordination costs) and the wage bill are given by

$$WB = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l). \quad (5)$$

However, organizations cannot simply be evaluated on the grounds of how cheaply they process information, but must also be able to swiftly react to new information. The notion of (the cost of) delay introduced in the previous paragraph can readily be generalized to multilayer hierarchies as follows:

$$C(\text{Delay}) = \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b. \quad (6)$$

As equation (6) makes clear, delay is the time it takes the hierarchy to process all information. In particular, this expression considers only one employee at each level of the hierarchy since managers at the same layer 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, the span of control and the communication efforts at each level of the hierarchy. Let  $\mathbf{c} = (c_1, \dots, c_{L-1})$  denote the vector of communication efforts exerted by the managers at the various layer of the hierarchy. Similarly, let  $\mathbf{n} = (n_1, \dots, n_{L-1})$ . The problem of the hierarchy is therefore

$$\min_{\mathbf{n}, \mathbf{c}, L} T(\mathbf{n}, \mathbf{c}, L) = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l) + \lambda \left( \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b \right) \quad (7)$$

for  $L \geq 2$ . As before,  $\lambda \in \mathbb{R}_{++}$  denotes the weight attached to delay (or, alternatively, the marginal cost of delay) relative to the wage bill. When only delay matters, I will often write  $\lambda = \infty$  to say that the weight attached to the wage bill,  $\omega$ , is zero. Of course, in that case the problem of the hierarchy reduces to minimizing (6) with respect to  $\mathbf{c}$ ,  $\mathbf{n}$  and  $L$ .

**Proposition 1** *A solution to programs (7) and (6) exist. Moreover, in the optimum  $\bar{n} \geq n_l^* \geq 2$  for all  $l = 1, \dots, L-1$ .*

---

<sup>15</sup> An important distinction in this model is between tasks and workloads. Workloads refer to the size of the task which is actually processed. Tasks, by contrast, refer to a manager's assignment, which could be delegated and therefore not necessarily processed. Of course, at the bottom level, tasks and workloads coincide.

Finally, I reiterate the fact that several restrictive assumptions implicit in (7) can easily be relaxed without affecting the conclusions of the paper. In particular, all the forthcoming results carry over to scenarios where

- The cost of delay is a strictly increasing function of delay, not necessarily linear.
- The time necessary to process one unit of information is an increasing function of a worker's workload,  $w_L$ . In particular, individual processing time is given by  $\eta(w_L)w_L$ , with  $\eta'(\cdot) \geq 0$ . Thus, the model can easily accommodate gains from specialization in information processing as modeled in Becker and Murphy (1992) or overload costs.
- Managers and production workers receive different hourly wages,  $\omega_m$  and  $\omega_p$ , instead of a common salary  $\omega$  per unit of time. In that case, the condition  $\lambda = \infty$  should be read as saying that both  $\omega_m$  and  $\omega_p$  are negligible compared to the weight attached to delay.

### 3 Optimal Communication in Uniform Hierarchies

To build some intuition for the results, this section focuses on the simpler case in which the span of control is constant across layers and equal to  $n$  (a uniform hierarchy). The analysis of the general case (7) is postponed to the next section.

For any given  $L \geq 3$  and  $n \geq 2$ , the problem of the (uniform) hierarchy is

$$\min_{\mathbf{c}} T(\mathbf{c}) = \eta N \prod_{i=1}^{L-1} D(n, c_i) + \sum_{l=1}^{L-1} n^{l-1} \tau(c_l) + \lambda \left( \eta N \prod_{i=1}^{L-1} \frac{D(n, c_i)}{n^{L-1}} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b \right). \quad (8)$$

The following two propositions characterize the optimal assignment of coordination responsibilities in uniform hierarchies.

**Proposition 2** *Let  $\lambda \in \mathbb{R}_{++}$  and  $\mathbf{c}^* = (c_1^*, c_2^*, \dots, c_{L-1}^*)$  be a vector of communication efforts that solves program (8). Then  $c_1^* \geq c_2^* \geq \dots \geq c_{L-1}^*$ .*

**Proof.** Assume instead that  $\tilde{\mathbf{c}} = (\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_{L-1})$ , with  $\tilde{c}_i < \tilde{c}_j$  for some  $i < j$ , solves (8). Then it is possible to improve upon  $\tilde{\mathbf{c}}$ . Indeed, let  $\check{\mathbf{c}}$  be a vector such that  $\check{c}_l = \tilde{c}_l$  for all  $l \neq i, j$  and  $\check{c}_i = \tilde{c}_j$  and  $\check{c}_j = \tilde{c}_i$ . Note that  $T(\tilde{\mathbf{c}}) - T(\check{\mathbf{c}}) = n^{i-1} \tau(\tilde{c}_i) + n^{j-1} \tau(\tilde{c}_j) - n^{i-1} \tau(\check{c}_j) - n^{j-1} \tau(\check{c}_i) = (n^{j-1} - n^{i-1}) [\tau(\tilde{c}_j) - \tau(\tilde{c}_i)] > 0$ . Thus  $\tilde{\mathbf{c}}$  cannot be optimal. This yields a contradiction. ■

Thus, in the optimum, senior managers work and are hence paid more than junior managers. Loss of control and lack of coordination also become relatively more severe as one goes down the hierarchy:  $D(n, c_{l-1}^*) \leq D(n, c_l^*)$  for all  $n$  and  $l$ .

**Proposition 3** *If  $\lambda = \infty$  and  $\mathbf{c}^*$  minimizes (6), (i) so does any permutation of  $\mathbf{c}^*$ . In particular,  $c_1^* = c_2^* = \dots = c_{L-1}^*$  whenever the solution is unique. (ii) Moreover, if  $-\frac{\tau'(c)}{D_c(n,c)}$  is increasing in  $c$  for all  $n$ ,<sup>16</sup> then  $c_1^* = c_2^* = \dots = c_{L-1}^*$ .*

The assumption that, for all  $n$ ,  $-\frac{\tau'(c)}{D_c(n,c)}$  is increasing in  $c$  is satisfied, for instance, whenever  $\tau''(\cdot) > 0$  and  $D_{cc}(\cdot, \cdot) = 0$  or  $\tau''(\cdot) = 0$  and  $D_{cc}(\cdot, \cdot) > 0$ . It postulates decreasing (net) returns to coordination and implies that coordination is more valuable when it is scarce. Indeed, in practice people typically deal with the most important issues immediately, and tie up details at a later stage. Functional forms consistent with all the assumptions on  $\tau(c)$  and  $D(n, c)$  are, for instance, either  $\tau(c) = \frac{c}{1-c}$  or  $\tau(c) = \alpha c^\beta$ ,  $\alpha > 0$ ,  $\beta > 1$ , and  $D(n, c) = n^{1-c}$ .

**Proof.** (i) When only delay matters, the problem of the hierarchy is

$$\min_{\mathbf{c}} \eta N \prod_{i=1}^{L-1} \frac{D(n, c_i)}{n^{L-1}} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b. \quad (9)$$

Claim (i) is then obvious since all coordination efforts enter this expression symmetrically.

(ii) Assume by contradiction that  $\tilde{\mathbf{c}} = (\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_{L-1})$  solves (9), with  $\tilde{c}_i > \tilde{c}_j$  for some  $i < j$  (the reasoning for  $\tilde{c}_i < \tilde{c}_j$  is similar). Let  $\mathbf{x} = (x_1, \dots, x_{L-1})$  with  $x_i = \frac{D(n, c_i)}{n}$ . Then (9) can be rewritten as  $\min_{x_1, \dots, x_{L-1}} \left\{ N\eta \prod_{l=1}^{L-1} x_l + \sum_{l=1}^{L-1} f(x_l) + (L-1)b \right\}$  where  $c_l = D^{-1}(n, nx_l)$  ( $D$  is strictly decreasing in  $c$  over the relevant domain  $n \leq \bar{n}$ ) and  $f(x_l) = \tau[D^{-1}(n, nx_l)]$ . Denote  $\tilde{x}_i = \frac{D(n, \tilde{c}_i)}{n}$ . Now consider a vector  $\check{\mathbf{x}}$  whose elements are the same as in  $\tilde{\mathbf{x}}$  except for the fact that  $\check{x}_i = k\tilde{x}_i$  and  $\check{x}_j = \frac{1}{k}\tilde{x}_j$ . Intuitively, starting at  $k = 1$ , as  $k$  increases, the difference between  $\check{x}_j$  and  $\check{x}_i$  gets smaller since by assumption  $\tilde{x}_i < \tilde{x}_j$ . Define  $g(k) = N\eta \prod_{l=1}^{L-1} \tilde{x}_l + \sum_{l \neq i, j}^{L-1} f(\tilde{x}_l) + f(k\tilde{x}_i) + f(\frac{1}{k}\tilde{x}_j) + (L-1)b$  to be the value of (9) at  $\check{\mathbf{x}}$  as a function of  $k$ . Clearly, if  $g'(1) < 0$ , then it is possible to improve upon  $\tilde{\mathbf{c}}$  and therefore  $\tilde{c}_i > \tilde{c}_j$  cannot be optimal. Note that  $g'(1) = f'(\tilde{x}_i)\tilde{x}_i - f'(\tilde{x}_j)\tilde{x}_j < 0$  provided  $f'(\tilde{x}_i) \leq f'(\tilde{x}_j)$  (since  $\tilde{x}_i < \tilde{x}_j$ ), that is, if  $f'(x)$  is increasing in  $x$ . By the Inverse Function Theorem, we get  $f'(x) = \frac{\tau'[D^{-1}(n, nx)]}{D_c(n, D^{-1}(n, nx))} = \frac{\tau'(c)}{D_c(n, c)}$ ,<sup>17</sup> which by assumption is decreasing in  $c$ . Claim (ii) then immediately follows from the fact that  $x = \frac{D(n, c)}{n}$  is decreasing in  $c$ . ■

**Example 1** Let  $D(n, c) = n^{1-c}$  and either  $\tau(c) = \frac{c}{1-c}$  or  $\tau(c) = \alpha c^\beta$ ,  $\alpha > 0$ ,  $\beta > 1$ , and assume interior solutions. Then program (8) is strictly convex and the unique optimal solution satisfies  $c_1^* > c_2^* > \dots > c_{L-1}^*$ . Furthermore,  $\partial \left( \frac{c_l^*}{c_{l+1}^*} \right) / \partial \lambda < 0$  and  $\lim_{\lambda \rightarrow \infty} \frac{c_l^*}{c_{l+1}^*} = 1$  for all  $l \leq L-2$ .

The intuition for Proposition 2 is simple: In the optimum higher-level managers coordinate more than lower-level managers since the former exert their influence on a much greater portion of the hierarchy than

<sup>16</sup>Throughout this paper when I say that something ‘is increasing’, ‘increases’ or ‘rises’, I mean to say that it is nondecreasing. Most of the inequalities in this paper are weak, so this convention leads to less awkwardness. When I want an inequality to be strict, I will say so explicitly, as in ‘strictly increasing’, etc.

<sup>17</sup>Note in fact that  $\tau(c)$  is by assumption a strictly increasing function and  $D^{-1}(n, nx)$  is strictly decreasing in  $x$  over the relevant domain  $2 \leq n \leq \bar{n}$ .



the latter do. Consider the wage bill in isolation. Although the various duplication functions (and hence the coordination efforts) 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, and that difference grows exponentially. Thus the actions of higher-level managers have a much greater impact on the performance of the organization than those taken by their subordinates.<sup>18</sup> By contrast, the cost of delay treats individual working time at each layer symmetrically since only the contribution of one manager at each layer has to be considered. The second part of Proposition 3, in particular, shows that when only delay matters, coordination efforts are equalized across layers. Since 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, except for the case where only delay matters.

Clearly, Proposition 2 provides a rationale for the empirical regularity 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 3 and Example 1, instead, may be useful to understand the recent trend towards empowerment. According to the model, in fact, a shift towards granting employees broader decision authority – on the way work is organized, in the present case – has to be expected when the weight attached to the cost of delay increases relative to the weight attached to the wage bill. Together, these results can be interpreted as formalizing the presence and indeed the optimality of overloads at the top levels of the hierarchy, overloads which however are suboptimal from a purely delay minimizing point of view.

The reorganization of Pepsi in 1988 provides a nice illustration of how demands for faster information processing driven by changes in the external business environment can lead to greater decentralization of the decision making process (Besanko et al., 2000). For Pepsi, a key change was the emergence of large regional supermarket chains. These in fact often operated in large territories, whereas Pepsi’s existing structures gave nobody region-wide authority over pricing. The problem was that regional executives, when faced with requests for promotions or special pricing deals, often disagreed over the appropriate strategy to follow. Their disputes had then to be resolved at the top levels of the hierarchy, which were forced to become involved in region-level pricing and promotion decisions. Not surprisingly, this impaired Pepsi’s ability to respond and put it at a competitive disadvantage in a market that demands nimble marketing responses to fast-changing circumstances. The solution was found by reorganizing the hierarchical chain of command in a new, geographically oriented matrix structure. This structure created the position of area general manager, who had final authority for operational decisions, such as pricing and promotions, within areas that were roughly the size of the territories of the large supermarket chains. Thus,

---

<sup>18</sup>Note that this is purely a cost-driven effect: coordination is just more cheaply provided near the top of the organization. In information processing hierarchies, however, it will be shown that not only are the costs of coordination lower near the top, but also the benefits may be larger.

the task of coordinating previously semiautonomous geographic areas was delegated down the hierarchy to an appositely created new managerial figure in response to an increase in the cost of delay driven by a change in the organization's business environment.<sup>19</sup>

The following proposition captures the intuition that, in large organizations, the top management is forced to spend a lot of time in coordination activities. Denote by  $c_1^*(L)$  the optimum value of  $c_1$  in a  $L$ -layer hierarchy.

**Proposition 4** *Consider program (8). Then  $c_1^*(L) \rightarrow 1$  as  $L \rightarrow \infty$ .*

Two remarks are in order. First, the fact that  $\lambda < \infty$  (which is implicit in (8)) is important. When only delay matters, in fact, it may well be the case that  $c_1^*(L)$  decreases in  $L$ . Second, and more importantly, note that when  $\tau(1) = \infty$  (which is true for instance if  $\tau(c) = \frac{c}{1-c}$ ), then increasing indefinitely the length of the hierarchy requires the top manager to bear arbitrarily large coordination costs. This model is therefore consistent with the view that the coordinating role of the top management is the main limiting factor of the size of the firm, and with Herbert Simon's claim that "attention is the chief bottleneck in organizational activities, and the bottleneck becomes narrower and narrower as we move to the tops of organizations, where parallel processing capacity becomes less easy to provide without damaging the coordinating function that is the prime responsibility of these levels" (Simon, 1976, p.294). Interestingly, the result holds despite the fact that coordination efforts are strategic substitutes (indeed, one can easily check that, by coordinating more, lower-level managers reduce the need for coordination at higher levels). Thus, Proposition 4 qualifies and extends the view that coordination limits the size of the firm because some tasks, most notably the coordinating function of the executive, cannot be delegated (see, e.g., Gifford (1992)).

## 4 Main Results: Communication vs. Span of Control

In this section I consider the general case in which the span of control can differ across layers and is chosen to minimize the organization's total costs. Starbuck, in particular, writes 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" (1971, p.88). This section studies how communication effort and the span of control interact to reduce duplications, and establishes conditions under which Starbuck's claim holds.

---

<sup>19</sup>Boeing is another case in point. According to J.R. 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" (p.191). As a result, instead of referring a problem upward in the hierarchy, coordination responsibilities were delegated to task forces and liaison groups of designers and engineers, who had to solve the problem at their own level, contacting and cooperating with peers in those departments affected by the new information.

Let the problem of the hierarchy be given by (7). This problem is considerably more complex than the previous one since now it involves  $2(L-1)+1$  endogenous variables,  $L$  of which can only take integer values. Nevertheless, a partial characterization of the optimal solution can still be obtained.

**Proposition 5** *For all  $L \geq 3$ , let  $(\mathbf{n}^*, \mathbf{c}^*) = ((n_1^*, c_1^*), \dots, (n_{L-1}^*, c_{L-1}^*))$  be a solution to (7). There are no  $k, h \geq 1$  such that  $c_k^* < c_{k+h}^*$  and  $n_k^* > n_{k+h}^*$ ,  $k+h \leq L-1$ . Moreover, if  $n_k^* = n_{k+h}^*$ , then  $c_k^* \geq c_{k+h}^*$ . If  $c_k^* = c_{k+h}^*$ , then  $n_k^* \leq n_{k+h}^*$ .*

Proposition 5 generalizes Proposition 2, for now it suffices to know that  $n_k^* \geq n_{k+h}^*$  to conclude that  $c_k^* \geq c_{k+h}^*$ . More importantly, Proposition 5 provides some support for the view that coordination in organizations is more cheaply provided when communication efforts are higher and spans of control are smaller at the top of the hierarchy. Indeed, the proof exploits the fact that, so long as  $c_k < c_{k+h}$  and  $n_k \geq n_{k+h}$  (or  $c_k \leq c_{k+h}$  and  $n_k > n_{k+h}$ ), one can always reduce the total wage cost of coordination  $\sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l)$  while keeping both delay and the workers' information processing workload constant by simply interchanging  $(n_k, c_k)$  with  $(n_{k+h}, c_{k+h})$ . However, the result falls short of showing that both small spans of control and more communication effort are always optimal near the top of the organization. Indeed, as the following proposition shows, to do so it is necessary to specify how communication effort and smaller spans of control interact to reduce duplications.

**Proposition 6** *If  $D(n, c)$  is log-supermodular,<sup>20</sup> then  $c_k^* \geq c_{k+h}^*$  and  $n_k^* \leq n_{k+h}^*$ .*

The requirement that  $D(n, c)$  be log-supermodular is very restrictive and in particular rules out the possibility that communication becomes more valuable as the size of a workgroup grows. On the other hand, it must also be stressed that this is only a sufficient condition, which becomes less and less stringent as the hierarchical gap  $h$  between two managers grows. For instance, it is easy to see that the wage cost of one additional hour of communication at layer  $l$  grows exponentially as one goes down the hierarchy:  $\prod_{i=0}^{l-1} n_i \geq 2^{l-1}$ . Thus, provided that the organization is concerned about wage costs and  $h$  is large ( $N$  and  $L$  are fixed), it is realistic to expect  $c_k^* \geq c_{k+h}^*$ , regardless of whether  $n_k^*$  is greater or smaller than  $n_{k+h}^*$ .<sup>21</sup> A simple example may help clarify the magnitude of this effect. Assume for instance that the managers at layer 1 and 6 of an organization spend the same time on communicating, and let the span of control of the managers at layer 1 be six, and the span of control of all the other managers be eight. Then the cost of a marginal increase in communication effort is almost 200,000 times smaller at layer 1 than it is at layer 6, since there are far more managers at the lower level. It is therefore likely that  $c_1^* \geq c_6^*$ , even if communication may be more 'effective' at the lower level, where the span of control is larger.

<sup>20</sup>Log-supermodularity of a positive function  $h(x, t)$  implies that the relative returns,  $h(x_H, t)/h(x_L, t)$ , are increasing in  $t$  for all  $x_H > x_L$ . (For a formal definition, see Athey (2002).) A simple example of a log-supermodular function is  $D(n, c) = v(n)d(c)$ .  $D(n, c) = n^{1-c}$  is log-submodular.

<sup>21</sup>A similar reasoning applies for the span of control.

Next, consider the problem of the hierarchy when only delay matters ( $\lambda = \infty$ ).

**Proposition 7** *Suppose (6) has a unique solution,<sup>22</sup> and denote it by  $(\mathbf{n}^*, \mathbf{c}^*)$ . Then  $c_1^* = c_2^* = \dots = c_{L-1}^*$  and  $n_1^* = n_2^* = \dots = n_{L-1}^*$ .*

Proposition 6 and 7 show that Keren and Levhari’s (1979) main result, that the span of control is decreasing as one travels up the hierarchy, with equality holding if wages are negligible relative to the marginal cost of delay, can be derived as a special case of this model. In addition, however, the present analysis also emphasizes possible complementarities between communication effort and the span of control in reducing duplications, as captured by the requirement that  $D(n, c)$  be not ‘too log-submodular’.

All in all, the above results seem in line with the evidence on the span of control. Table 3.6 in Galbraith (1977, p.48), for instance, shows that in the production departments of United States and Canadian oil refineries, the span decreases as one passes from second level (Foreman) to the third (General Foreman), and then again from the third to the fourth (Superintendent). Interestingly, Keren and Levhari (1979) also argue that, in military organizations, where the cost of staffing the hierarchy is secondary compared to the utility of planning time saved, spans of control tend to be more uniform. However, these shreds of evidence are far from conclusive and different models, in particular Beggs (2001) and Calvo and Wellisz (1979), have found the opposite result, namely that top managers control more direct subordinates than junior managers do.

## 4.1 Communicative Skills

So far this paper has focused on situations in which all managers are equally talented. I showed that higher-level managers will 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, since all our managers are identical, it does not really matter which hierarchical position a manager is allocated to.

However, when differences in ability are taken into account, this indifference result is no longer likely to be valid, and several models have in fact shown that the more able agents should be assigned to the higher levels of the hierarchy. Unfortunately, these models by and large focus on talent as either information processing speed (e.g. Prat (1997)) or as the ability to monitor subordinates (e.g. Calvo and Wellisz (1979)). Casual empiricism, on the other hand, suggests that the ability to coordinate effectively

---

<sup>22</sup>Due to integer constraints, it is difficult to solve program (6) and to show that a solution is unique. However, ignoring such constraints, one can assume that (6) 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. Moreover, if  $D(n, c)/n$  is decreasing in  $n$  over the relevant range, then  $n_1^* = \dots = n_{L-1}^* = \bar{n}$ .

is very important at the top levels of the hierarchy, probably more important than the ability to process information rapidly. Indeed, it is often required that successful candidates for executive positions have ‘strong interpersonal and communicative skills, a collegial management style’, ‘ability to organize work and demonstrated ability to work in harmony with people’, ‘ability to communicate effectively’.<sup>23</sup> Within the economics literature, Williamson emphasizes the same point when he writes:

The bounds of rationality here take the form of language rather than computational limits and evidently vary among individuals. If the specialization of labor is feasible, those whose rationality limits are less severely constrained than others are natural candidates to assume technical, administrative or political leadership positions – which is to say that hierarchy can emerge on this account (1975, p.24).

This subsection presents a simple variant of the problem of the hierarchy which endogenizes the choice of a manager’s communicative skills, or ability. To keep things simple, assume that every manager spends the same amount of time  $a$  coordinating the work of his direct subordinates ( $a$  may be interpreted as the length of a normal working day). However, managers differ in their communicative skills. Specifically, let  $c \in [0, 1]$  index communicative skills and  $\tau(c)$  be the wage paid to a manager of ability  $c$ . The fact that  $\tau$  is increasing in  $c$  therefore simply means that higher-ability managers are paid more than their lower-ability counterparts. Furthermore, let  $D(n, c)$  measures the amount of duplications that arise during the delegation process when a superior’s ability is  $c$  and  $n$  is the size of the workgroup. Obviously, the abler the superior, the fewer the duplications. The modified problem of the hierarchy is

$$\min_{\mathbf{n}, \mathbf{c}, L} T(\mathbf{n}, \mathbf{c}, L) = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l) + \lambda \left( \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + (L-1)(a+b) \right). \quad (10)$$

Note that the only difference between program (10) and (7) is that the time cost of coordinating subordinates at any given layer  $l$  is now  $a + b$  and not  $\tau(c_l) + b$ . Notwithstanding this, it is not difficult to see that the results of the previous sections carry over to the new framework. In particular, it is clear that, in line with Williamson’s claim, managers with strong interpersonal and communicative skills will typically be found at the top echelons of the hierarchy. In addition, however, the fact that the time cost of coordinating subordinates is now independent of  $c$  yields a novel implication. When only delay matters, in fact, there is now no tradeoff between fewer duplications and higher wages, and therefore in the optimum  $c_1^* = \dots = c_{L-1}^* = 1$ .<sup>24</sup> Thus, empirically, the result suggests that, as urgency increases, the average quality of the management should increase, especially at the lower and middle levels of the hierarchy. In

---

<sup>23</sup>The Bureau of the Census (1995) also showed that, when recruiting new production staff, US employees ranked communication skills above, e.g., previous work experience, recommendations from current employees, years of schooling and grades. According to the employees, the only factor in making hiring decisions which was more important than communication skills was “applicant’s attitude”.

<sup>24</sup>Similarly, if  $D(n, c)/n$  is decreasing in  $n$  over the relevant range, then  $n_1^* = \dots = n_{L-1}^* = \bar{n}$  is also optimal.

particular, the model makes the testable prediction 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, especially at the bottom of the hierarchy, than companies operating in mature sectors.

## 5 The Flattening Hierarchy

This section focuses on relationship between the optimal number of layers and the span of control, and relates changes in these elements of formal structure of the organization to changes in the relative importance of reducing delay versus increasing specialization. My goal is to use the present model to investigate the widespread belief among practitioners that ‘delaying’ (i.e., the reduction the number of formal layers of an organization) is a powerful instrument to speed up decision making. Jeffrey Immelt, the CEO of General Electric, for instance, 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."<sup>25</sup> Consistent with this statement, Rajan and Wulf also show that in recent years US corporations have experienced both a reduction in the number of formal layers and an increase in the number of managers directly reporting to the top management.

In this section, I consider a stylized model incorporating both a concern for delay and gains from specialization in information processing. With the help of this model, I demonstrate that if, in a sense made precise below, delegation is mainly driven by specialization (rather than delay), as urgency increases, the organization becomes flatter and the span of control increases. Using a specific information processing technology, I also show that this condition is more likely to be fulfilled the greater the fixed delay  $b$  that delegation involves and the strength of the gains from specialization, and the smaller the coordination costs.

To focus on the span of control and the number of levels in the hierarchy, a very simple version of the basic model is considered. The following three assumptions are made. First, communication effort is assumed to take only two possible values,  $c \in \{0, 1\}$ , with  $\tau(1) = \bar{\tau}$  and  $\tau(0) = 0$ . Second, the duplication function is given by  $D(n, c) = n^{1-c}$  if  $n \leq \bar{n}$  and  $D(n, c) = n$  otherwise. Third, the span of control is constant across layers. The first and the second assumption together guarantee that in the optimum all managers exert high effort (i.e.,  $c_l^* = 1$  for all  $l \leq L-1$ ),<sup>26</sup> but of course any  $D(n, 0)$  sufficiently high would

<sup>25</sup>Rajan and Wulf (2004) p.2. They 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.

<sup>26</sup>This claim follows immediately from the fact that, if  $c = 0$ , then  $D(n, 0) = n$  and there are costs but no benefits from delegation. For completeness, I prove this claim in the appendix (Lemma A1), where I also show that in the optimum  $\bar{n} \geq n^* \geq 2$ .

suffice. More generally, in fact, the qualitative results of this section hold provided  $D(n, 0)$  is sufficiently high and  $D(n, 1)$  is sufficiently small over the relevant domain  $n \leq \bar{n}$ .

A crucial implication of the above assumptions is that the hierarchy can be fully characterized by only two variables,  $n$  and  $L$ . This is important because it allows one to focus exclusively on their interaction between  $n$  and  $L$ , without having to worry, for instance, about the impact of a change in the span of control at one layer on the span at another layer. Generalizing the model to consider interactions between variables at different layers would certainly be interesting but would also greatly complicate the analysis and is therefore left for future research.

The additional key ingredient of this model is the presence of positive returns from specialization in information processing. Most of the information processing literature, in fact, with the important exception of Bolton and Dewatripont (1994), assumes that delegation is driven exclusively by the need to reduce delay. However, that has the unfortunate consequence that the number of levels in the organization will tend to increase with urgency. This section illustrates how that effect can be offset when specialization is introduced into the model.

Gains from specialization are modeled as follows. Let  $\eta(w_L)$  be the time it takes a worker to process one unit of information.  $\eta(w_L)$  is assumed to be increasing and convex in individual workload,  $w_L$ . Thus, as in Becker and Murphy (1992), smaller individual workloads and hence greater specialization allow workers to process each unit of information faster. Note that, realistically, returns from specialization are assumed to be decreasing.

Knowing that  $c_l^* = 1$  for all  $l \leq L - 1$ , the problem of the hierarchy with specialization can be written as

$$\min_{n, L} T^S(n, L; \lambda) = \underbrace{N\eta\left(\frac{N}{n^{L-1}}\right) + \bar{\tau} \sum_{l=1}^{L-1} n^{l-1}}_{WB(n, L)} + \underbrace{\lambda \left[ \frac{N}{n^{L-1}} \eta\left(\frac{N}{n^{L-1}}\right) + (L-1)(\bar{\tau} + b) \right]}_{DL(n, L)} \quad (11)$$

for all  $\bar{n} \geq n \geq 2$  and  $L \geq 2$ . A key property of (11) is the following:

**Proposition 8** *Assume  $L \geq 2$ . Then  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$ .*

Quasi-supermodularity captures a weak notion of complementarity (see Milgrom and Shannon (1994) for a formal definition) and provides a precise meaning to the intuition that the span of control and the number of levels are substitutes. Indeed, their role is very similar in this model, as they both allow the organization to increase parallel information processing by expanding its size. Note that, perhaps intuitively, the result does not hold when  $L = 1$ ; in that case, in fact, if the organization wishes to increase the span of control, it must also necessarily create a new hierarchical level, and  $n$  and  $L$  are therefore complementary. This extreme case can be ruled out by assuming, for instance, that  $\eta N \lambda > 2[\bar{\tau} + \lambda(\bar{\tau} + b)]$ , which guarantees that the one-worker hierarchy is dominated by a two-layer hierarchy with two subordinates.

Program (11) differs in one major respect compared to (7), the crucial difference being that the new problem of the hierarchy when only wages matter ( $\lambda = 0$ ) is non-trivial. In fact, note that multilayer hierarchies can now be optimal even when minimizing the wage bill is the only concern of the hierarchy designer, whereas under program (7) this could have never been the case. (One may therefore say that delegation in (7) is driven *exclusively* by the need to reduce delay.) To state the main result of this section, however, more definitions are needed. Let  $L_W^*(n)$  be the number of layers that minimize the wage bill,  $WB(n, L)$ , given  $n$ . Similarly, let  $L_D^*(n)$  and  $L^*(n)$  be, respectively the number of layers that minimize delay ( $DL(n, L)$ ) and total costs ( $T^S(n, L)$ ), given  $n$ . Note that, for fixed  $n$ , both the wage bill and delay (and hence  $T^S$ ) are strictly convex in  $L$ ,<sup>27</sup> and therefore all these numbers are essentially unique.<sup>28</sup> I say that delegation is *mainly* driven by specialization (delay), and write  $L_W^* \geq L_D^*$  ( $L_D^* \geq L_W^*$ ), if  $L_W^*(n) \geq L_D^*(n)$  ( $L_D^*(n) \geq L_W^*(n)$ ) for all  $\bar{n} \geq n \geq 2$ .

Clearly, if delegation is mainly driven by specialization, as reducing delay becomes more important one may expect hierarchies to become flatter and spans of control to increase (the second intuition is motivated by the fact that, as far as delay is concerned, larger spans are always beneficial). Furthermore, these two effects would reinforce each other since  $n$  and  $L$  are substitutes. The next Proposition formalizes this intuition.

**Proposition 9** *Suppose that  $L \geq 2$  and  $L_W^* \geq L_D^*$ . Then the solution  $(n^*, -L^*)$  to (11) is increasing in the strong set order in  $\lambda$ .<sup>29</sup>*

Proposition 9 relies on the fact that one can restrict attention to  $L_W^* \geq L \geq L_D^*$  since any  $L$  outside this range cannot be optimal. Since we already know that  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$ , it is only necessary to show that the single-crossing property in  $(n, -L; \lambda)$  is satisfied over the relevant range (the assumption that  $L_W^* \geq L_D^*$  is essential here). The result then immediately follows from Theorem 4 in Milgrom and Shannon (1994).

However, a problem with this result is that it is not clear when one can reasonably expect delegation to be mainly driven by specialization ( $L_W^* \geq L_D^*$ ). Inspection of (11) suggests that this condition is more likely to be satisfied when  $\bar{\tau}$  is low and  $b$  high. Furthermore, one might also expect this condition to hold when returns to specialization are substantial. The following example confirms all these intuitions.

**Example 2** Suppose that  $\eta(w) = \beta w^\alpha$ , with  $\beta > 0$ ,  $\alpha \geq 1$ . Then a sufficient condition for  $L_W^* \geq L_D^*$  to hold is that  $b \geq \frac{1}{2(2^\alpha - 1)} \bar{\tau}$ .

<sup>27</sup>To be precise,  $T^S(n, L)$  is said to be strictly convex if, for all  $L$ ,  $T^S(n, L+2) - T^S(n, L+1) > T^S(n, L+1) - T^S(n, L)$ , and similarly for  $WB(n, L)$  and  $DL(n, L)$ .

<sup>28</sup>The only complication may arise from the fact that  $L$  is a discrete variable. However, "ties" between values of  $L$  are non-generic and would disappear if, for instance,  $\bar{\tau}$  was slightly perturbed. That is, the set of parameters  $\bar{\tau}$  for which  $T^S(n, L'; \lambda) = T^S(n, L''; \lambda)$ ,  $L', L'' \in \arg \min_L T^S(n, L; \lambda)$ , has Lebesgue measure zero in  $\mathbb{R}$ .

<sup>29</sup>For a definition of the strong set order, see Milgrom and Shannon (1994).



The condition  $b \geq \frac{1}{2(2^\alpha - 1)} \bar{\tau}$  is more likely to be satisfied when  $b$  and  $\alpha$  are high (note that the latter measures the strength of the gains from specialization), and  $\bar{\tau}$  is low. The proof of example 2 (which is included in the Appendix) also shows that  $b > 0$  is a necessary condition for  $L_W^* \geq L_D^*$ . Thus, in my framework, delegation can be mainly driven by specialization only if there is some substantial source of delay in the communication process which is not accounted for in the wage bill (however, note that  $L_W^* \geq L_D^*$  is only a sufficient condition for the result to hold).

## 6 Information Processing Hierarchies

This section focuses on a generalization of the production model which allows for the reporting of information to one's direct superior. The scenario I have in mind is a situation where information must initially be processed by employees at the bottom of the organization (perhaps because they are the ones in direct contact with the customers) and then transmitted up the hierarchy through formal communication channels. The contribution of this section, besides checking the robustness of the main results of the paper, will be to provide a second rationale for the fact that superiors spend more time on coordination activities than their subordinates.

In information processing hierarchies, all the information is eventually transmitted to the top manager through reports which summarize and filter it. The basic model in Section 2 is extended as follows. As before, upper-level managers coordinate the work of their direct subordinates and bottom-level employees process raw information. However, now each subordinate is also assumed to send a report to his immediate superior, the size of which is proportional to the information he has processed. Formally, a manager at layer  $l > 1$  processing an amount of information  $w_l$  will send a report of size  $R_l = rw_l$  to his direct superior, where  $r \geq 0$  is a parameter measuring information compression. Thus, since at layer  $L$  the workload of a bottom-level manager is  $N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ , he will send a report of size  $R_L = rN \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ . Managers at layer  $L - 1$  receive their subordinates' reports  $n_{L-1}R_L = rN \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ , process all the 'relevant' information (see below) and send their own reports up the hierarchy until all the information has been communicated the top manager. Note that  $r$  must be sufficiently less than one because otherwise the organization would be better off if the superior processed that information alone.

Two important feature of this framework must be pointed out. The first feature is that skip-level reporting is not allowed: subordinates can only report to their direct superiors. This clearly involves some loss of generality since we know from Radner (1993) that skip-level reporting is an important feature of optimal information processing hierarchies. However, since this practice seems much less common in actual organizations than the information processing literature would suggest, it is probably useful to consider, at least as a first step, a scenario in which skip-level reporting is not allowed.

A second important feature of the model is that reports will in general contain duplications made at lower levels. For instance, the discussion above made clear that the reports received at layer  $L - 1$  contain duplications made at layer  $L$ . My analysis will distinguish two polar cases: the no-detection and the detection scenario. In the latter case, managers will be assumed to be able to detect and disregard duplications made at lower levels at no cost. In the former scenario, instead, this will be impossible and superiors will process duplications. Furthermore, their reports will also contain duplications made at lower levels. Needless to say, in reality we would expect something in-between these two polar scenarios. Finally, it must be stressed that both scenarios are generalizations of the basic model of Section 2 to the case where  $r \neq 0$ .

## 6.1 No-Detection Scenario

In the no-detection case, managers process and summarize the information they receive from their subordinates in a non-selective way. Thus, for example, since  $R_l = rw_l$ , the workload of a manager at layer  $L - 1$  is  $w_{L-1} = rN \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ . More generally, we have that

$$w_l = n_l R_{l+1} = n_l r w_{l+1} \quad (12)$$

for  $l = 1, \dots, L - 1$  or, equivalently

$$w_l = r^{L-l} N \prod_{i=1}^{L-1} D(n_i, c_i) \left[ \prod_{i=0}^{l-1} n_i \right]^{-1} \quad (13)$$

for all  $l = 1, \dots, L$ . Clearly, equations (12) and (13) imply that higher-level managers process duplications made at lower levels, and also that their reports contain duplications. For all  $L \geq 2$ , the problem of the hierarchy becomes

$$\min_{\mathbf{n}, \mathbf{c}} \underbrace{\sum_{l=1}^L \left\{ \eta N r^{L-l} \prod_{i=1}^{L-1} D(n_i, c_i) + \prod_{i=0}^{l-1} n_i \tau(c_l) \right\}}_{\text{Wage Bill}} + \lambda \underbrace{\sum_{l=1}^L \left( \eta N r^{L-l} \prod_{i=1}^{L-1} D(n_i, c_i) \left[ \prod_{i=0}^{l-1} n_i \right]^{-1} + \tau(c_l) \right)}_{C(\text{Delay})} \quad (14)$$

where of course  $c_L = 0$  since bottom-level employees do not coordinate the work of other managers. Note that program (8) is a special case of (14) as the latter reduces to the former when  $r = 0$ . However, since  $\prod_{i=1}^{L-1} D(n_i, c_i)$  still appears at all the levels of the hierarchy, it is easy to see that all the results of sections 3 and 4 carry over to the new scenario. In particular

**Proposition 10** *Let the span of control be constant across layers and  $\mathbf{c}^* = (c_1^*, c_2^*, \dots, c_{L-1}^*)$  solve program (14). If  $\lambda \in \mathbb{R}_{++}$ , then  $c_1^* \geq c_2^* \geq \dots \geq c_{L-1}^*$ . If  $\lambda = \infty$  and  $-\frac{\tau'(c)}{D_c(n, c)}$  is increasing in  $c$  for all  $n$ , then  $c_1^* = c_2^* = \dots = c_{L-1}^*$ .<sup>30</sup>*

<sup>30</sup>The proof is analogous to those of Proposition 2 and 3 and is therefore omitted.

## 6.2 Detection Scenario

In this scenario, superiors are assumed to be able to skip duplicated information at no cost (that is, without processing it) and therefore their reports contain no duplications. If superiors immediately detect duplicated tasks and do not process them twice, individual workloads are given by

$$w_l = r^{L-l} N \prod_{i=1}^{\min\{l, L-1\}} D(n_i, c_i) \left[ \prod_{i=0}^{l-1} n_i \right]^{-1} \quad (15)$$

for all  $l = 1, \dots, L$ . The problem of the hierarchy therefore becomes

$$\min_{\mathbf{n}, \mathbf{c}} \sum_{l=1}^L \left\{ \eta N r^{L-l} \prod_{i=1}^{\min\{l, L-1\}} D(n_i, c_i) + \prod_{i=0}^{l-1} n_i \tau(c_l) \right\} + \lambda \sum_{l=1}^L \left( \eta N r^{L-l} \prod_{i=1}^{\min\{l, L-1\}} D(n_i, c_i) \left[ \prod_{i=0}^{l-1} n_i \right]^{-1} + \tau(c_l) \right) \quad (16)$$

for  $\lambda \in \mathbb{R}_{++}$  and

$$\min_{\mathbf{n}, \mathbf{c}} \sum_{l=1}^L \left( \eta N r^{L-l} \prod_{i=1}^{\min\{l, L-1\}} D(n_i, c_i) \left[ \prod_{i=0}^{l-1} n_i \right]^{-1} + \tau(c_l) \right) \quad (17)$$

for  $\lambda = \infty$ . The crucial difference between (16)-(17) and all the programs studied above is that now there is an asymmetry between duplications made at upper levels and those made at the lower levels, since the former affect a greater number of levels than the latter do. In particular, it should be clear that the organization now has an additional incentive to reduce duplications at the top than it had in the production or in the no-detecting scenarios. Thus, if, as it is intuitive, duplications increase as the span of control gets larger, higher-level managers will typically spend more time communicating and have more limited spans of control than their subordinates:

**Proposition 11** *Suppose  $D(n, c) \leq D(n+1, c)$  for all  $n$  and  $c$ . Let  $L \geq 3$  and  $(\mathbf{n}^*, \mathbf{c}^*)$  be a solution to (16). Then there are no  $k, h \geq 1$  such that  $c_k^* < c_{k+h}^*$  and  $n_k^* > n_{k+h}^*$ ,  $k+h \leq L-1$ . Moreover, if  $D(n, c)$  is log-supermodular, then  $c_k^* \geq c_{k+h}^*$  and  $n_k^* \leq n_{k+h}^*$ .*

Note that, as in Proposition 6, a log-supermodularity assumption must be imposed on the duplication function to ensure that in the optimum  $c_k \geq c_{k+h}$  and  $n_k \leq n_{k+h}$ . This condition is however now even less restrictive because of the greater incentives to invest in communication effort and reduce the span of control near the top of the hierarchy. These incentives can perhaps be best appreciated in uniform hierarchies (the proof is analogous to that of Proposition 2 and is therefore omitted).

**Proposition 12** *Suppose the span of control is constant across layers and  $\mathbf{c}^* = (c_1^*, c_2^*, \dots, c_{L-1}^*)$  solves either program (16) or (17). Then  $c_1^* \geq c_2^* \geq \dots \geq c_{L-1}^*$ .*

Relative to production hierarchies and the no-detection case, senior managers now spend more time on coordinating than their subordinates not only because it is less costly (as it was before), but also because their effort has an impact on a greater number of levels. Indeed, (15) makes clear that a manager's

communication effort only affects the workloads of the managers below him. Thus, when comparing the effects of an increase in, say,  $c_1$  and  $c_3$ , we must take into account that there are layers (in this case layers 2 and 3) in which an increase in  $c_1$  reduces processing time but an increase in  $c_2$  does not. This asymmetry is most evident when only delay matters. In fact, in production hierarchies, Proposition 3 guarantees that under some conditions communication efforts are equalized across layers. Now, in contrast, under the same conditions it may well be the case that  $c_1^* > \dots > c_{L-1}^*$ , for instance if solutions are interior and  $D(n, c) = n^{1-c}$  and  $\tau(c) = \frac{c}{1-c}$ . Of course, this does not imply that empowerment will not occur, as it is still true that senior managers exert more effort than their subordinates for cost reasons, and that effect disappears as minimizing delay becomes more and more important. Thus, the same forces at work in the production scenario also operate in information processing hierarchies.

## 7 Final Remarks

Coordination is an essential ingredient for survival and success in competitive environments, and ways to enhance it are a central concern for modern management. Liaison groups, teamwork, the use of intranets are just a few examples of working practices that more and more firms utilize to create and promote ‘horizontal linkages’ between interdependent organizational units, and constitute an important source of competitive advantage for companies such as Du Pont, Merck and Cisco. This paper has made a first attempt to provide a characterization of the optimal assignment of coordination responsibilities in a hierarchical organization. Nevertheless, many important issues have been neglected.

One major omission is that issues related to the optimal provision of incentives have been assumed away by imposing that all agents share the same objective. However, incentive problems have already received a lot of attention in the literature, so I will say no more here. A second limitation of this paper, which I have already mentioned, is that tasks are simply assumed to be divided evenly among subordinates. This assumption is very convenient in that it allows me to restrict attention to balanced hierarchies, which are both tractable and quite realistic, at least as a first approximation. However, fully endogenizing the division of tasks among subordinates would certainly be desirable. A third assumption that may be worth relaxing is the fact that, in my framework, tasks are homogeneous across the levels of the hierarchy. This may be done, for instance, by parametrizing the duplication functions in a way that reflects the greater complexity of the job as one goes up the hierarchy. In particular, in that scenario, I would expect the result that senior managers coordinate more than their subordinates to be strengthened, provided of course that the marginal returns to coordination rise with complexity.

Finally, I would like to emphasize that duplication of effort is certainly not the only problem that poor communication brings about. In real organizations, in fact, it is also often the case that some tasks are not carried out (‘information loss’) or that they are carried out in a way that makes it difficult to integrate

them with other tasks ('poor integration'). In the remainder of this section, I will try to argue that these issues can, at least to a certain extent, be incorporated in the present framework.

Consider loss of useful information first. Many authors have noticed that the greater the number of levels in a system, the greater will be the probability of "noise" in functional connections among organizations units (e.g., Williamson (1967), Starbuck (1971)). Assume therefore that the exact specification of the tasks delegated to subordinates is subject to random shocks, or "misunderstandings". In particular, suppose that with some probability these misunderstandings prevent some of the items (or, in a continuous model, a segment of positive measure) from being processed. Clearly, if the cost of the resulting loss of information is high enough, it might be optimal to specify partially overlapping tasks so that the risk that a particular piece of information is not processed is minimized. Similarly, consider a scenario where the firm's employees (say 1 and 2) must integrate their tasks, X and Y, with another task, Z. Ideally, only one of the workers should be put in charge of Z (say 1); communication between 1 and 2 would then ensure that Y and Z are properly coordinated. In practice, however, the integration of Z with Y would probably be more difficult than with X, perhaps because of intrinsic difficulties in articulating tacit knowledge or more simply because of a lack of incentives to help. Clearly, a radical way to solve the problem would be to assign task Z to both agents, so that each of them would be responsible for an integrated task,  $X+Z$  or  $Y+Z$ .

The discussion above clearly suggests that when the costs of information loss or poor integration are very high, a certain amount of redundancy in task assignments may indeed be optimal and thus indicates two possible ways in which the duplication function may be micro-founded. The present model could then be reinterpreted by saying that communication reduces misunderstandings. All results would straightforwardly carry over to the new scenario, and reductions in amount of 'optimal' duplications would provide a simple way to quantify the benefits of fewer misunderstandings.

# Appendix: Omitted Proofs

## Proof of Proposition 1

Clearly, in the optimum  $L < \infty$ , otherwise delay would diverge. To show that in the optimum  $\bar{n} \geq n_l \geq 2$  for all  $l$ , assume instead that either  $n_l = 1$  or  $n_l > \bar{n}$  for some  $l$ . In either case, the size of a manager's task at layer  $l$  and  $l+1$  is the same, and therefore no reduction in delay accrues. Clearly the organization would be better off by eliminating layer  $l$ , thus saving at least the fixed communication cost  $b$ . Hence, neither  $n_l = 1$  nor  $n_l > \bar{n}$  can be optimal. The existence of a solution for the communication efforts follows from Weierstrass theorem. Thus a solution to programs (7) and (6) exists. ■

## Proof of Example 1

I first show that, if  $D(n, c) = n^{1-c}$  and  $\tau(c) = \frac{c}{1-c}$ , then  $c_l^* > c_{l+1}^*$  and  $\partial \left( \frac{c_l^*}{c_{l+1}^*} \right) / \partial \lambda < 0$ ,  $l = 1, \dots, L-2$ . Then I show that (8) is strictly convex. The case  $\tau(c) = \alpha c^\beta$ ,  $\alpha > 0$ ,  $\beta > 1$  is similar and therefore omitted.

Assuming interior solutions, the first order conditions with respect to  $c_k$  and  $c_{k+1}$  yield  $(n^{k-1} + \lambda) (1 - c_k^*)^{-2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i^*}$  and  $(n^k + \lambda) (1 - c_{k+1}^*)^{-2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i^*}$ . Clearly,  $c_k^* > c_{k+1}^*$  (so long as  $n > 1$ , of course). Furthermore

$$\frac{1 - c_k^*}{1 - c_{k+1}^*} = \left( \frac{n^k + \lambda}{n^{k-1} + \lambda} \right)^{-1/2},$$

which implies that  $\lim_{\lambda \rightarrow \infty} \frac{c_l^*}{c_{l+1}^*} = 1$  and that  $\frac{1-c_k^*}{1-c_{k+1}^*}$  is strictly increasing in  $\lambda$ . Moreover,  $\frac{\partial c_k^* / \partial \lambda}{\partial c_{k+1}^* / \partial \lambda} < \frac{1-c_k^*}{1-c_{k+1}^*}$ . Note that  $\frac{\partial c_k^* / \partial \lambda}{\partial c_{k+1}^* / \partial \lambda} < \frac{c_k^*}{c_{k+1}^*}$  implies that  $\partial \left( \frac{c_k^*}{c_{k+1}^*} \right) / \partial \lambda < 0$ . Since  $\frac{c_k^*}{c_{k+1}^*} > 1 > \frac{1-c_k^*}{1-c_{k+1}^*}$ , then  $\partial \left( \frac{c_k^*}{c_{k+1}^*} \right) / \partial \lambda < 0$ .

To show that  $T(\mathbf{c})$  is strictly convex in  $\mathbf{c}$ , compute its Hessian matrix  $H_{T(\mathbf{c})} = \left[ \frac{\partial^2 T(\mathbf{c})}{\partial c_i \partial c_j} \right]_{i,j=1,\dots,L-1}$  where  $\frac{\partial^2 T(\mathbf{c})}{\partial c_i \partial c_j} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i} (\ln(n))^2 \equiv a > 0 \ \forall i, j, i \neq j$  and  $\frac{\partial^2 T(\mathbf{c})}{\partial c_k^2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i} (\ln(n))^2 + (n^{k-1} + \lambda) (1 - c_k)^{-2} \equiv a + b_k > 0$ .

By applying elementary row and column operations, the Hessian matrix can be rewritten as

$$\tilde{H}_{T(\mathbf{c})} = \begin{bmatrix} b_1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & b_k & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & b_{L-2} & 0 \\ x_1 & x_2 & \cdots & \cdots & \cdots & x_{L-2} & x_{L-1} + b_{L-1} \end{bmatrix}$$

where  $x_k = a \sum_{i=1}^k \frac{b_i}{b_i} > 0$ ,  $k = 1, \dots, L-1$ . Since the determinant of a lower-triangular matrix is simply the product of its diagonal entries,  $\det(H_{T(\mathbf{c})}) = \det(\tilde{H}_{T(\mathbf{c})}) > 0$ . The same reasoning can also be applied to

all the leading principal minors of  $H_{T(\mathbf{c})}$ . Consequently,  $H_{T(\mathbf{c})}$  is positive definite at every point,  $T(\mathbf{c})$  is strictly convex in  $\mathbf{c}$  and vector  $\mathbf{c}^*$  that solve the first order conditions is the unique global minimum. ■

#### Proof of Proposition 4

First consider corner solutions. If  $c_1^*(L) = 1$  there is nothing to prove. Clearly,  $c_1^*(L) = 0$  cannot be optimal for large  $L$  because  $\prod_{i=1}^{L-1} D(n, c_i)$  would then diverge. Thus we are left with interior solutions. By differentiating (8) with respect to  $c_1$ , compute the marginal benefit ( $MB$ ) and the marginal cost ( $MC$ ) of an increase in  $c_1$  in the optimum for given  $L$ :

$$\left. \frac{\partial T(\mathbf{c})}{\partial c_1} \right|_{\mathbf{c}=\mathbf{c}^*} = - \underbrace{\left[ -D_c(n, c_1^*) \left( 1 + \frac{\lambda}{n^{L-1}} \right) \eta N \prod_{i=2}^{L-1} D(n, c_i^*) \right]}_{MB(L)} + \underbrace{(1 + \lambda) \tau'(c_1^*)}_{MC(L)}.$$

Clearly, if  $MB(L) \rightarrow \infty$  as  $L \rightarrow \infty$ , then  $c_1^*(L) \rightarrow 1$  since  $\tau'(\cdot) > 0$ . Assume instead that  $c_1^*(L)$  does not converge to 1 as  $L$  grows large. Then there exists a sequence  $\{L_k\}_{k=1}^\infty$  such that  $|c_1^*(L_k) - 1| \geq \delta > 0$  for all  $k$ . Since the domain of  $c$ ,  $[0, 1]$ , is compact, we can choose this sequence so that  $c_1^*(L_k) \rightarrow \hat{c}$ ,  $|\hat{c} - 1| \geq \delta$  and hence  $D(n, \hat{c}) > 1$ . Note that  $MB(L) \geq -D_c(n, c_1^*) \left( 1 + \frac{\lambda}{n^{L-1}} \right) \eta N [D(n, c_1^*)]^{L-2}$ . However, this last inequality implies that  $\lim_{k \rightarrow \infty} MB(L_k) = \infty$ , a contradiction. ■

#### Proof of Proposition 5

Suppose not. In particular, let  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}}) = ((\tilde{n}_1, \tilde{c}_1), \dots, (\tilde{n}_{L-1}, \tilde{c}_{L-1}))$  be a solution to (7) and assume that  $\tilde{c}_k < \tilde{c}_{k+h}$  and  $\tilde{n}_k \geq \tilde{n}_{k+h}$  for some  $k$  and  $h$ ,  $1 \leq k < k+h \leq L-1$  (the case where  $\tilde{c}_k \leq \tilde{c}_{k+h}$  and  $\tilde{n}_k > \tilde{n}_{k+h}$  is similar). Consider a vector  $(\check{\mathbf{n}}, \check{\mathbf{c}})$  whose elements are the same as in  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  except that  $(\tilde{n}_k, \tilde{c}_k)$  and  $(\tilde{n}_{k+h}, \tilde{c}_{k+h})$  have been interchanged. That is,

$$(\check{\mathbf{n}}, \check{\mathbf{c}}) = (\dots, (\tilde{n}_{k-1}, \tilde{c}_{k-1}), (\tilde{n}_{k+h}, \tilde{c}_{k+h}), (\tilde{n}_{k+1}, \tilde{c}_{k+1}), \dots, (\tilde{n}_{k+h-1}, \tilde{c}_{k+h-1}), (\tilde{n}_k, \tilde{c}_k), (\tilde{n}_{k+h+1}, \tilde{c}_{k+h+1}), \dots).$$

Note that

$$\eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \lambda \left( \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) \right) \Big|_{(\mathbf{n}, \mathbf{c})=(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})} = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \lambda \left( \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) \right) \Big|_{(\mathbf{n}, \mathbf{c})=(\check{\mathbf{n}}, \check{\mathbf{c}})}.$$

However  $\sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l) \Big|_{(\mathbf{n}, \mathbf{c})=(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})} > \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l) \Big|_{(\mathbf{n}, \mathbf{c})=(\check{\mathbf{n}}, \check{\mathbf{c}})}$ . In fact, the only terms in  $\sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l)$  that matter for the comparison are the ones from  $k$  to  $k+h$  included. For every  $k < l < k+h$  we have  $(\tilde{n}_k - \tilde{n}_{k+h}) \prod_{i=0, i \neq k}^{l-1} \tilde{n}_i \tau(\tilde{c}_l) \geq 0$ . Moreover,  $\tau(\tilde{c}_k) + \prod_{i=k}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) > \tau(\tilde{c}_{k+h}) + \prod_{i=k+1}^{k+h} \tilde{n}_i \tau(\tilde{c}_k)$  since

$$\left( \tilde{n}_k \prod_{i=k+1}^{k+h-1} \tilde{n}_i - 1 \right) \tau(\tilde{c}_{k+h}) > \left( \tilde{n}_{k+h} \prod_{i=k+1}^{k+h-1} \tilde{n}_i - 1 \right) \tau(\tilde{c}_k).$$

Thus  $T(\tilde{\mathbf{n}}, \tilde{\mathbf{c}}, L) > T(\check{\mathbf{n}}, \check{\mathbf{c}}, L)$ ,  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  cannot be optimal and this yields a contradiction. ■

### Proof of Proposition 6

The proof proceeds by contradiction. Let  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  be a candidate solution and consider the following three cases:

- (a)  $\tilde{c}_k < \tilde{c}_{k+h}$  and  $\tilde{n}_k > \tilde{n}_{k+h}$ . This can never be optimal by Proposition 5.
- (b)  $\tilde{c}_k < \tilde{c}_{k+h}$  and  $\tilde{n}_k \leq \tilde{n}_{k+h}$ . Let  $(\check{\mathbf{n}}, \check{\mathbf{c}}) = (\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  except that  $\check{c}_k = \tilde{c}_{k+h}$  and  $\check{c}_{k+h} = \tilde{c}_k$ . Say that  $(\check{\mathbf{n}}, \check{\mathbf{c}})$  improves upon  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  if  $T(\tilde{\mathbf{n}}, \tilde{\mathbf{c}}, L) > T(\check{\mathbf{n}}, \check{\mathbf{c}}, L)$ . This is clearly the case whenever

$$A[D(\tilde{n}_k, \tilde{c}_k)D(\tilde{n}_{k+h}, \tilde{c}_{k+h})] \geq A[D(\tilde{n}_k, \tilde{c}_{k+h})D(\tilde{n}_{k+h}, \tilde{c}_k)] \quad (18)$$

and

$$\prod_{i=0}^{k-1} \tilde{n}_i \tau(\tilde{c}_k) + \prod_{i=0}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) > \prod_{i=0}^{k-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) + \prod_{i=0}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_k) \quad (19)$$

where  $A = \eta N \left[ 1 + \lambda \left( \prod_{i=1}^{L-1} \tilde{n}_i \right)^{-1} \right] \prod_{i \neq k, k+h} D(\tilde{n}_i, \tilde{c}_i)$ . Under our assumptions, condition (19) is satisfied. Condition (18) is satisfied for all  $\tilde{c}_k < \tilde{c}_{k+h}$  and  $\tilde{n}_k \leq \tilde{n}_{k+h}$  if  $D(n, c)$  is log-supermodular.

- (c)  $\tilde{c}_k \geq \tilde{c}_{k+h}$  and  $\tilde{n}_k > \tilde{n}_{k+h}$ . The proof is analogous to that of part b). Let  $(\check{\mathbf{n}}, \check{\mathbf{c}}) = (\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  except that  $\check{n}_k = \tilde{n}_{k+h}$  and  $\check{n}_{k+h} = \tilde{n}_k$ . A sufficient condition for  $(\check{\mathbf{n}}, \check{\mathbf{c}})$  to improve upon  $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$  is that

$$A[D(\tilde{n}_k, \tilde{c}_k)D(\tilde{n}_{k+h}, \tilde{c}_{k+h})] \geq A[D(\tilde{n}_k, \tilde{c}_{k+h})D(\tilde{n}_{k+h}, \tilde{c}_k)] \quad (20)$$

where  $A$  was defined above and

$$\tilde{n}_k \sum_{l=k}^{k+h} \prod_{\substack{i=0 \\ i \neq k}}^{l-1} \tilde{n}_i \tau(\tilde{c}_l) > \tilde{n}_{k+h} \sum_{l=k}^{k+h} \prod_{\substack{i=0 \\ i \neq k}}^{l-1} \tilde{n}_i \tau(\tilde{c}_l). \quad (21)$$

Condition (21) is satisfied under our assumptions. Condition (20) is satisfied for all  $\tilde{c}_k \geq \tilde{c}_{k+h}$  and  $\tilde{n}_k > \tilde{n}_{k+h}$  if  $D(n, c)$  is log-supermodular.

Together, (a), (b) and (c) imply that log-supermodularity of  $D(n, c)$  is sufficient to guarantee that in the optimum  $c_k \geq c_{k+h}$  and  $n_k \leq n_{k+h}$ . ■

### Proof of Proposition 7

Obvious from the symmetry of equation (6). ■

The following lemma proves a claim in Section 5.

**Lemma A1** Consider program (7). Under the assumptions made in Section 5 (including those on  $\eta(\cdot)$ ),  $\bar{n} \geq n^* \geq 2$  and  $c_l^* = 1$  for all  $l \leq L-1$  and  $L \geq 2$ .

**Proof.** Suppose not. In particular, suppose that  $c_l^* = 0$ . This implies that the size of an individual task at layer  $l$  and  $l+1$  are the same, since  $D(n, 0) = n$ . But then it is optimal to eliminate layer  $l$ , to avoid



at the very least the fixed cost  $b$ . Thus  $c_l^* = 0$  cannot be optimal. That  $\bar{n} \geq n^* \geq 2$  follows immediately from Propositions 1. ■

To show that  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$  for all  $L \geq 2$ , the following result is useful.

**Lemma A2** Let  $f : X \times Y \rightarrow \mathbb{R}$ ,  $X = \{\underline{x}, \underline{x} + 1, \dots, \bar{x} - 1, \bar{x}\}$ ,  $Y = \{\underline{y}, \underline{y} + 1, \dots, \bar{y} - 1, \bar{y}\}$ ,  $\bar{x}, \underline{y} \in \mathbb{N}$ . If

$$f(x + 1, y + 1) + f(x, y) \geq f(x + 1, y) + f(x, y + 1)$$

for all  $(x, y) \in X \times Y$ , then  $f$  is supermodular in  $(x, y)$  on  $X \times Y$ .<sup>31</sup>

**Proof.** Rearranging the inequality in Lemma A2 yields

$$f(x + 1, y) - f(x, y) \leq f(x + 1, y + 1) - f(x, y + 1) \leq \dots \leq f(x + 1, y + z) - f(x, y + z) \leq \dots \quad (22)$$

and similarly

$$f(x, y + 1) - f(x, y) \leq f(x + 1, y + 1) - f(x + 1, y) \leq \dots \leq f(x + d, y + 1) - f(x + d, y) \leq \dots \quad (23)$$

for some  $(x + d, y + z) \in X \times Y$ . From (23), it follows that  $f(x + d, y) - f(x, y) \leq f(x + d, y + 1) - f(x, y + 1)$ . Recursion (22) can then be applied. After some iterations, we get  $f(x + d, y) - f(x, y) \leq f(x + d, y + z) - f(x, y + z)$ . Rearranging this last expression yields

$$f(x + d, y) + f(x, y + z) \leq f(x + d, y + z) + f(x, y),$$

which proves that  $f$  is indeed supermodular in  $(x, y)$ . ■

### Proof of Proposition 8

In order to prove that  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$  for all  $L \geq 2$  (and of course  $n \geq 2$ ), I first show that  $\frac{N}{n^{L-1}}$  is supermodular in  $(n, -L)$  for all  $L \geq 2$ . By Lemma A2, we only need to check that

$$\begin{aligned} \frac{1}{(n+1)^L} + \frac{1}{n^{L-1}} &\geq \frac{1}{n^L} + \frac{1}{(n+1)^{L-1}} \\ \iff \left(\frac{n+1}{n}\right)^L &\geq \left(\frac{n}{n-1}\right) \\ \iff L &\geq \frac{\ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{n+1}{n}\right)} \equiv q(n). \end{aligned}$$

Note that  $1 < q(n) < 2$  for  $n \geq 2$ . In fact, both  $\frac{n}{n-1} > \frac{n+1}{n}$  and  $q < 2 \Leftrightarrow n^2 - n - 1 > 0$  are verified for all  $n \geq 2$ . (Alternatively, one can verify that  $q(n)$  is decreasing in  $n$  and that  $\ln(2)/\ln(3/2) \simeq 1.709$ .) Thus,  $L \geq 2$  is a sufficient condition for  $\frac{N}{n^{L-1}}$  to be supermodular in  $(n, -L)$ .

---

<sup>31</sup>Lemma A2 can be thought of as variant for integer-valued choice variables of the ‘marginal’ rule stating that, for a twice continuously differentiable  $f : X \rightarrow \mathbb{R}$ ,  $X \subset \mathbb{R}^l$ , supermodularity is equivalent to checking that  $\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0$  for all  $x \in X$  and all distinct  $i, j = 1, \dots, l$ .

Now, individual working time  $\frac{N}{n^{L-1}}\eta\left(\frac{N}{n^{L-1}}\right)$  is a *strictly* increasing function of a supermodular function, and hence quasi-supermodular. It immediately follows that delay is quasi-supermodular in  $(n, -L)$  for  $L \geq 2$ . To see that the same is true for the wage bill, note that

$$\bar{\tau} \sum_{l=1}^L (n+1)^{l-1} + \bar{\tau} \sum_{l=1}^{L-1} n^{l-1} \geq \bar{\tau} \sum_{l=1}^L n^{l-1} + \bar{\tau} \sum_{l=1}^{L-1} (n+1)^{l-1} \iff (n+1)^{L-1} \geq n^{L-1}, \quad (24)$$

which obviously is always verified. Note the tricky logic here:  $\eta\left(\frac{N}{n^{L-1}}\right)$  is not necessarily quasi-supermodular since we only required  $\eta(\cdot)$  to be *weakly* increasing; however, the wage bill is quasi-supermodular because of (24). ■

### Proof of Proposition 9

In this proof I first show that, with no loss of generality, it is possible to restrict attention to  $L_W^* \geq L \geq L_D^*$ , as any  $L$  outside this range will not be optimal. Then I show that  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$  and satisfies the single-crossing property in  $(n, -L; \lambda)$  over the relevant range. The result then follows from Theorem 4 in Milgrom and Shannon (1994).

The first thing to notice is that, for any given  $n$ ,  $\eta'' \geq 0$  ensures that the wage bill and delay (and hence  $T^S$ ) in (11) are strictly convex in  $L$  (see also footnote 27). This implies that  $L_W^*$ ,  $L_D^*$  and  $L^*$  are all essentially unique (see footnote 28) and therefore we can restrict attention to  $L_W^* \geq L \geq L_D^*$ . To see the last point, fix  $n = n^*$  and note that  $L > L_W^*$  cannot be a solution to (11) because  $L_W^*$  would reduce both the wage bill (by assumption) and delay (since  $L > L_W^* \geq L_D^*$  and delay is convex), relative to  $L$ . The case where  $L < L_D^*$  is similar. Furthermore,  $-T^S(n, L; \lambda)$  is quasi-supermodular in  $(n, -L)$ , provided that  $L \geq 2$ . In fact, the proof of Proposition 8 shows that  $\frac{N}{n^{L-1}}$  is supermodular in  $(n, -L)$ , and it is well-known that strictly increasing transformations of supermodular functions are quasi-supermodular. Thus, it only remains to be shown that  $-T^S(n, L; \lambda)$  satisfies the single-crossing property in  $(n, -L; \lambda)$  over the relevant range. That is, we need to show that, for all  $\lambda'' > \lambda' > 0$ , if

$$-\lambda'' \left[ \frac{N}{n''^{L''-1}} \eta \left( \frac{N}{n''^{L''-1}} \right) + (L'' - 1)(\bar{\tau} + b) \right] > 0, \text{ then } -\lambda' \left[ \frac{N}{n'^{L'-1}} \eta \left( \frac{N}{n'^{L'-1}} \right) + (L' - 1)(\bar{\tau} + b) \right] > 0$$

for all  $(n'', -L'') > (n', -L')$  in the set  $\{(n, L) \mid 2 \leq n \leq \bar{n}, L \geq 2, L_W^*(n) \geq L \geq L_D^*(n)\}$ . This is true if and only if

$$\frac{N}{n''^{L''-1}} \eta \left( \frac{N}{n''^{L''-1}} \right) + (L'' - 1)(\bar{\tau} + b) < \frac{N}{n'^{L'-1}} \eta \left( \frac{N}{n'^{L'-1}} \right) + (L' - 1)(\bar{\tau} + b). \quad (25)$$

Obviously, larger spans of control always make (25) more likely to be satisfied. A decrease in  $L$  also reduces delay since, for all  $n$  (and in particular  $n''$ ),  $L_W^*(n) \geq L \geq L_D^*(n)$  and delay is convex in  $L$ . ■

### Proof of Example 2

Fix  $n$ . Let  $L_W$ ,  $L_D$  be, respectively, the number of layers that minimize the wage bill and delay given  $n$ . (For notational ease, asterisks denoting optimal values and the reference to  $n$  will be omitted in the

following.) Due to the convexity of the wage bill,  $L_W$  is the lowest value of  $L$  that satisfies

$$WB(n, L) \leq WB(n, L+1) \Leftrightarrow N\eta\left(\frac{N}{n^{L-1}}\right) + \bar{\tau} \sum_{l=1}^{L-1} n^{l-1} \leq N\eta\left(\frac{N}{n^L}\right) + \bar{\tau} \sum_{l=1}^L n^{l-1}$$

or, equivalently,

$$N \left[ \eta\left(\frac{N}{n^{L-1}}\right) - \eta\left(\frac{N}{n^L}\right) \right] \leq \bar{\tau} n^{L-1}. \quad (26)$$

Similarly,  $L_D$  is the lowest value of  $L$  that satisfies

$$\frac{N}{n^{L-1}} \left[ \eta\left(\frac{N}{n^{L-1}}\right) - \frac{1}{n} \eta\left(\frac{N}{n^L}\right) \right] \leq \bar{\tau} + b. \quad (27)$$

At  $L = L_W$ , (27) can be written as

$$\underbrace{N \left[ \eta\left(\frac{N}{n^{L_W-1}}\right) - \eta\left(\frac{N}{n^{L_W}}\right) \right] - \bar{\tau} n^{L_W-1}}_{\leq 0} + N \frac{n-1}{n} \eta\left(\frac{N}{n^{L_W}}\right) \leq b n^{L_W-1}.$$

Thus, a sufficient condition for  $L_W \geq L_D$  is that

$$b \geq (n-1) \frac{N}{n^{L_W}} \eta\left(\frac{N}{n^{L_W}}\right). \quad (28)$$

This condition however depends on  $L_W$ . To obtain an explicit solution, assume that  $\eta(w) = \beta w^\alpha$ , with  $\beta > 0$  and  $\alpha \geq 1$ . (26) then becomes  $N \left[ \beta \left(\frac{N}{n^{L-1}}\right)^\alpha - \beta \left(\frac{N}{n^L}\right)^\alpha \right] \leq \bar{\tau} n^{L-1}$ . Simple algebra yields  $n^L \geq N \left[ (n^\alpha - 1) \frac{\beta n}{\bar{\tau}} \right]^{\frac{1}{1+\alpha}}$ . Since  $L_W$  is the lowest value of  $L$  that satisfies this equation, we have that

$$n^{L_W} \geq N \left[ (n^\alpha - 1) \frac{\beta n}{\bar{\tau}} \right]^{\frac{1}{1+\alpha}}. \quad (29)$$

Clearly  $N \left[ (n^\alpha - 1) \frac{\beta n}{\bar{\tau}} \right]^{\frac{1}{1+\alpha}}$  is a lower bound to  $n^{L_W}$ . Plugging this value into the right-hand side of (28) gives us the following sufficient condition for  $L_W \geq L_D$ :

$$b \geq \frac{(n-1)}{(n^\alpha - 1)n} \bar{\tau}.$$

The inequality in example 2 then easily follows from the fact the condition  $L_W \geq L_D$  must hold for all  $n$  and the fact the right-hand side of the above equation is decreasing in  $n$ . ■

### Proof of Proposition 11

The proof is similar to those of Proposition 5 and 6 and only a sketch is given. It is useful to divide total information processing costs (wages and delay) into three parts, namely those incurred a) at layers 1 to  $k-1$ , b) at layers  $k$  to  $k+h-1$  and c) at all the remaining levels. Consider a putative solution where  $n_k > n_{k+h}$  and  $c_k < c_{k+h}$ .

- a) Neither  $(n_k, c_k)$  nor  $(n_{k+h}, c_{k+h})$  has any impact on information processing costs at layers 1 to  $k-1$ . Thus swapping  $(n_k, c_k)$  for  $(n_{k+h}, c_{k+h})$  has no impact on (16).
- b) Only  $(n_k, c_k)$  has an impact on the information processing costs at layers  $k$  to  $k+h-1$ . Clearly, it is beneficial to swap  $(n_k, c_k)$  for  $(n_{k+h}, c_{k+h})$ , since duplications are decreasing in  $n$  (and  $c_k < c_{k+h}$ ).
- c) Both  $(n_k, c_k)$  and  $(n_{k+h}, c_{k+h})$  have an impact on information processing costs (and of course on total coordination costs). The same arguments used to prove Proposition 5 show that it is beneficial to swap  $(n_k, c_k)$  for  $(n_{k+h}, c_{k+h})$ .

It is therefore possible to improve on the putative solution. This yields a contradiction.

The proof of the second part of the proposition is analogous to the proof of Proposition 6. ■

## References

- [1] ATHEY, S., (2002), “Monotone Comparative Statics Under Uncertainty”, *Quarterly Journal of Economics*, **117**, 187-223.
- [2] BECKER, G.S. and MURPHY, K.M. (1992), “The Division of Labor, Coordination Costs and Knowledge”, *Quarterly Journal of Economics*, **107**, 1137-1160.
- [3] BEGGS, A.W. (2001), “Queues and Hierarchies”, *Review of Economic Studies*, **68**, 297-322.
- [4] BESANKO, D., DRANOVE, D. and SHANLEY, M. (2000), *Economics of Strategy* (New York: John Wiley & Sons).
- [5] BOLTON, P. and DEWATRIPONT, M., (1994), “The Firm as a Communication Network”, *Quarterly Journal of Economics*, **109**, 809-839.
- [6] BOLTON, P. and FARRELL J. (1990), “Decentralization, Duplication, and Delay”, *Journal of Political Economy*, **98**, 803-826.
- [7] BUREAU OF THE CENSUS (1995), “First Findings from the EQW National Employer Survey”, EQW Catalog Number: RE01.
- [8] CALVO, C.A. and WELLISZ, S. (1978), “Supervision, Loss of Control and the Optimal Size of the Firm”, *Journal of Political Economy*, **86**, 943-952.
- [9] CALVO, C.A. and WELLISZ, S. (1979), “Hierarchy, Ability and Income Distribution”, *Journal of Political Economy*, **87**, 991-1010.
- [10] CHANDLER, A.D. (1990), *Scale and Scope: the Dynamics of Industrial Capitalism* (Cambridge, Massachusetts: Harvard University Press).
- [11] CHWE, M.S.Y. (1995), “Strategic Reliability of Communication Networks” (mimeo).
- [12] DESSEIN, W. and SANTOS, T. (2003), “The Demand for Coordination” (Working Paper, Chicago Graduate Business School).
- [13] GALBRAITH, J.R. (1977), *Organization Design* (Reading, Massachusetts: Addison-Wesley Publishing Company).
- [14] GIFFORD, S. (1992), “Allocation of Entrepreneurial Attention”, *Journal of Economic Behavior and Organization*, **19**, 265-284.
- [15] GUETZKOW, H. (1981), “Communications in Organizations”, in Nystrom P.C. and Starbuck W.H. (ed.), *Handbook of Organizational Design* (Oxford: Oxford University Press).

- [16] HARRIS, M. and RAVIV, A. (2002), “Organization Design”, *Management Science*, **48**, 852-865
- [17] HART, O. and MOORE, J. (1999), “On the Design of Hierarchies: Coordination Versus Specialization” (Harvard Institute for Economic Research, Discussion Paper No. 1880).
- [18] JEHIEL, P. (1999), “Information Aggregation and Communication in Organizations”, *Management Science*, **45**, 659-669.
- [19] KEREN, M. and LEVHARI, D. (1979), “The Optimum Span of Control in a Pure Hierarchy”, *Management Science*, **25**, 1162-1172.
- [20] KEREN, M. and LEVHARI, D. (1983), “The Internal Organization of Firms and the Shape of Average Costs”, *Bell Journal of Economics*, **14**, 474-486.
- [21] KEREN, M. and LEVHARI, D. (1989), “Decentralization, Aggregation, Control Loss and Costs in a Hierarchical Model of the Firm”, *Journal of Economic Behavior and Organization*, **11**, 213-236.
- [22] MARIN, D. and VERDIER, T. (2003), “Globalization and the New Enterprise”, *Journal of the European Economic Association, Papers and Proceedings*, **1**, 337-344.
- [23] MAHONEY T.A., JERDEE T.H and CARROLL S.J. (1965), “The Job(s) of Management”, *Industrial Relations*, **4**, 97-110.
- [24] MEAGHER, K.J. (2003), “Generalizing Incentives and Loss of Control in an Optimal Hierarchy: the Role of Information Technology”, *Economic Letters*, **78**, 273-280.
- [25] MILGROM, and SHANNON, C. (1994), “Monotone Comparative Statics”, *Econometrica*, **62**, 1109-1146.
- [26] PRAT A. (1997), “Hierarchies of Processors with Endogenous Capacity”, *Journal of Economic Theory*, **77**, 214-222.
- [27] RADNER, R. (1993), “The Organization of Decentralized Information Processing”, *Econometrica*, **61**, 1109-1146.
- [28] RAJAN, R.G. and WULF, J. (2004), “The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies” (Working Paper, Chicago Graduate Business School).
- [29] RICE, S.A. (1940), “The Role and Management of the Federal Statistical System”, *American Political Science Review*, **34**, 481-488.
- [30] ROSS, S.M. (1983), *Introduction to Stochastic Dynamic Programming* (New York, Academic Press).

- [31] ROTEMBERG, J.J. (1999), “Process- versus Function-Based Hierarchies”, *Journal of Economics & Management Strategy*, **8**, 453-487.
- [32] SAYLES, L.R. (1964), *Managerial Behavior* (New York: McGraw-Hill).
- [33] SIMON, H.A. (1976), *Administrative Behavior* (New York: Free Press).
- [34] SIMON, H.A., SMITHBURG D.W. and THOMPSON V.A. (1950), *Public Administration* (New York: Knopf).
- [35] STARBUCK, W.H. (1971), “Organization Growth and Development”, 11-141, in STARBUCK, W.H. (ed.), *Organizational Growth and Development* (Harmondsworth: Penguin Books).
- [36] URWICK, F.W. (1956), “The Manager’s Span of Control”, *Harvard Business Review*, **34**, 39-47.
- [37] VAN ZANDT, T. (1995), “Continuous Approximations in the Study of Hierarchies”, *Rand Journal of Economics*, **26**, 575-590.
- [38] VAN ZANDT, T. (2004), “Balancedness of Real-Time Hierarchical Resource Allocation” (CEPR Discussion Paper No. 4276).
- [39] VAYANOS, D. (2003), “The Decentralization of Information Processing in the Presence of Interactions”, *Review of Economic Studies*, **70**, 667-695.
- [40] WILLIAMSON, O.E. (1967), “Hierarchical Control and Optimum Firm Size”, *Journal of Political Economy*, **75**, 123-138.
- [41] WILLIAMSON, O.E. (1975), *Markets and Hierarchies: Analysis and Antitrust Implications* (New York, Free Press).