Contract Complexity, Incentives, and the Value of Delegation

Nahum Melumad
Graduate School of Business
Columbia University
New York, NY 10027

Dilip Mookherjee
Boston University
Boston, MA 02215

Stefan Reichelstein
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and
BWZ
University of Vienna
Vienna, Austria

In settings where the revelation principle applies, delegation arrangements are frequently inferior to centralized decision making, and at best achieve the same level of performance. This paper studies the value of delegation when organizations are constrained by a bound on the number of contingencies in any contract. For a principal-agent setting with asymmetric information, we compare centralized mechanisms where the principal retains sole responsibility for contracting and coordinating production, with delegation mechanisms where one agent (a manager) is delegated authority to contract with other agents and coordinate production. Relative to centralization, delegation entails a control loss, but allows decisions to be more sensitive to the manager's private information. We identify circumstances under which the flexibility gain outweighs the control loss, so that delegation emerges superior to centralized contracting.

Mookherjee and Reichelstein gratefully acknowledge financial support from National Science Foundation grant SES 9209455.

Journal of Economics & Management Strategy, Volume 6, Number 2, Summer 1997, 257–289

1. Introduction

The Revelation Principle has played a central role in the theory of incentives. It presumes that agents can costlessly enter into comprehensive contracts and process unlimited amounts of information. While the Revelation Principle has proven useful in understanding the incentive constraints imposed by private information, it has also been an impediment to the theory of organization design. As Myerson (1982) demonstrated, when this principle applies, a completely centralized organization performs at least as well as any other organizational arrangement. In particular, any noncooperative equilibrium outcome of a decentralized organization can be replicated by a centralized revelation mechanism where all agents communicate their private information to a center and receive instructions from it concerning actions to be taken. It thus precludes a theory successful in explaining the widespread prevalence of decentralized decision making in organizations and contracting situations.

Optimal revelation mechanisms provide a useful performance benchmark for evaluating particular organizational arrangements observed in practice. For instance, Myerson (1981) and Myerson and Satterthwaite (1983) have used the characterization of optimal revelation mechanisms to examine the relative efficiency of specific auction and bargaining procedures. Similarly Baron and Besanko (1992), Gilbert and Riordan (1995), McAfee and McMillan (1995), and Melumad et al. (1995) have examined the performance of delegated contracting arrangements, using optimal revelation mechanisms as the reference point.[1] Nevertheless, as long as the Revelation Principle applies, one cannot explain why centralized revelation mechanisms are not more widely used in practice.

Our approach in this paper is to dispense with one of the more questionable assumptions underlying the Revelation Principle. Specifically, we postulate that writing detailed incentive contracts is costly; in particular, contracts corresponding to revelation mechanisms are prohibitively costly. Under this hypothesis, we compare centralized decision making with a decentralized scheme in which the principal communicates and contracts with only one agent (the *manager*) and

1. Similar perspectives have been adopted in the literature on regulation (e.g. Laffont and Tirole, 1993) and intrafirm resource allocation (e.g., Harris et al., 1982; Kanodia, 1993).

delegates authority to that agent to communicate and contract with other agent(s).[2]

We view a contract as a collection of "if . . . , then . . ." statements that specify the parties' obligations (the "then . . ." parts) for different contingencies (the "if . . ." parts). Depending on the organizational context, a contingency will be determined either by agents' reports, or by their action choices, which are publicly verifiable. Since each contingency has to be specified in advance as part of the formal contract, often with the aid of lawyers, it is plausible to suppose that, everything else remaining the same, contracts with more contingencies are more expensive to write and understand.[3] Moreover, in case of a contract dispute, it will take third parties (such as courts) more time and effort to comprehend contracts involving more contingencies.[4]

Our model does not specify an explicit cost for including contingencies. We only require that contracts corresponding to revelation mechanisms would be prohibitively costly, since they typically involve an infinite number of contingencies. As a consequence, contracts will be incomplete in the sense that they are not fully state-contingent.[5] To

2. This is a broader notion of delegation than one in which the principal retains authority over contracting with all agents but allows the latter to decide their own production levels. Issues concerning such narrower forms of delegation can be raised in a single-agent setting. It turns out that limiting the number of contract contingencies does not serve to distinguish such forms of delegation from centralization, a point discussed in further detail following Proposition 2. It is for this reason that we consider the broader notion of delegation in this paper.

3. Hart and Holmstrom (1987) have argued that in some contexts the costs of incorporating a continuum of contingencies into a contract may not be prohibitively large. For instance, if the parties can specify a simple mathematical function for the relationship between contingencies and decisions taken by the principal, the complexity of the contract will not vary significantly between contexts where the domain of this function is binary, has a finite number of elements, or forms a continuum. In certain contexts this may indeed be the case, e.g., where contracts are indexed proportionally to the rate of inflation. However, in many other contexts contingencies are defined by a multitude of variables (e.g. measures of quantity or quality of outputs delivered, cost conditions, or the state of technology), whose relationship to subsequent production assignments and transfer payments cannot be represented by a simple mathematical formula. In such situations it seems plausible that the contract has to be phrased as a list of "if . . . , then . . ." statements, whence the length of the list will affect the costs of writing the contract.

4. For an analysis of costly contract contingencies in a general equilibrium setting see Dye (1985).

5. In the literature on *incomplete contracts* the basic assumption is that the state of the world is not verifiable even though it is known to the contracting parties. Moreover, parties make specific investments that are subject to a holdup problem. Parts of this literature have considered the use of revelation mechanisms to provide the parties with appropriate incentives for specific investment; see, for example, Green and Laffont (1988), Rogerson (1992), and Che and Hausch (1996). While our analysis does not include specific investments, it emphasizes asymmetric information and imposes contractual incompleteness for complexity reasons, much in the spirit of Williamson (1975, 1985).

keep the analysis tractable, our model assumes that each agent's private information is real-valued, and that contracts can involve only a finite number of contingencies. As a consequence the information held privately by the agents is rich in comparison with what they can report to others, i.e., the set of contingencies that can be formally incorporated in the contract.[6] In addition, if payments are conditioned on actions chosen by agents, then only a finite number of action choices are "permitted," in order to limit the complexity of the contract.

When contracts cannot be fully state-contingent, delegated contracting gains a potential advantage over centralization. Under delegation, the principal empowers one agent (the manager) to decide the agents' production assignments. These decisions are made on the basis of better information than the principal could possibly obtain under centralization, owing to the limited nature of reports submitted by agents. We shall refer to this effect as the *flexibility gain* inherent in delegated contracting. On the other hand, earlier work on contractual hierarchies has shown that delegated contracting is prone to experience a *control loss*: the manager (respectively, the prime contractor) has an inherent tendency to exploit his monopsony power to procure too little from subordinates (the subcontractors). This is also known in the vertical-integration literature as the distortion arising from the *double marginalization of rents*. Our earlier work [Melumad, Mookherjee, and Reichelstein (MMR hereafter), 1995] has shown that absent any restriction on contracts, the principal can overcome the control loss by monitoring external procurement and subsidizing it at a suitable rate. Under certain additional assumptions regarding the sequence of communication and contracts, delegation can then be shown to replicate the performance of the optimal revelation mechanism.

With limitations on the number of contract contingencies it will be impossible for the principal to completely eliminate the control loss that accompanies delegated contracting. As a consequence, the performance comparison between centralized and delegated contracting depends on the relative magnitude of the flexibility gain and the control loss. Our main result in this paper is that, irrespective of the number of admissible contract contingencies, the flexibility gain *always* out-

6. Thus, a more realistic formulation would represent the private information of each agent as involving a large number of dimensions, all of which cannot be formally represented in the contract. We shall not pursue this approach, because the analysis of optimal contracts would become significantly more difficult.

weighs the control loss, thus implying that delegated contracting is the preferred organizational mode.[7]

This result requires a number of conditions that include risk neutrality, an appropriate sequencing of contracts, the absence of collusion between the manager and his subordinates, and, perhaps most important, the principal's ability to monitor the manager's production contribution.[8] By example we demonstrate that if the principal were to observe only the aggregate team output, the resulting control loss under delegation could become sufficiently severe so as to reverse our ranking and render centralization superior. The importance of the other requirements in determining the performance of delegation was explained in MMR (1995) in the context of costless contracting. In the current context it is evident that in the absence of one or several of the above requirements, delegation may perform worse than centralization.[9]

In earlier work (MMR, 1992) we have also investigated the costs and benefits of delegated contracting, but in a setting of costly communication. Specifically, agents were constrained to report messages from a finite set, while contracts could still be based on a continuum of possible action contingencies. Since action choices can serve as a partial substitute for communication of reports, delegated contracting was less constrained in MMR (1992). In the setting of this paper, the limit on contract contingencies implies restrictions on both reports and action choices, which additionally constrains delegated contracting compared to the setting where only communication is restricted. Hence the results obtained in this paper are stronger than those obtained previously. In addition, in this paper we obtain sufficient conditions for delegated contracting to *strictly* dominate centralization, whereas the earlier paper only established conditions for weak dominance.

The paper is organized as follows. The model and the corresponding optimal revelation mechanism are presented in Section 2. The optimal centralized contract subject to limited contingencies is developed in the first part of Section 3. In Proposition 1, we show that because of

7. Other agency models with limited communication include Green and Laffont (1986, 1987) and Laffont and Martimort (1996).

8. Ability to monitor financial payments to subordinates (respectively, subcontractors) would also suffice, if monitoring of production contributions were not possible, for the reasons explained in MMR (1992).

9. For instance, with a sufficiently large number of contingencies, the performance levels of either contracting mode will be close to those obtained in a context with an unlimited number of contingencies. Since centralized contracting achieves superior performance when contracting is costless and any of the above requirements is not met, it follows from a continuity argument that centralization will also be superior when the number of contingencies allowed in the contract is large enough.

the constraint on contracts the principal prefers that the agents report sequentially rather than simultaneously (as in a revelation mechanism). Proposition 2 then demonstrates that given the optimal centralized contracting arrangement with sequential reporting (corresponding to any given number of contingencies), there exists a delegated contracting scheme involving fewer contingencies, which generates a level of expected profit for the principal that is at least as high. Under some additional assumptions, Proposition 3 shows that profits generated by delegation are strictly higher. Concluding remarks are contained in Section 4.

2. The Model

We consider a team production problem involving a principal and two agents.[10] Agent i can take some productive action a_i from a set A_i. The set of jointly feasible actions is the product $A = A_1 \times A_2$. For every action profile $a \equiv (a_1, a_2)$, the principal receives a monetary benefit $B(a)$. If agent i takes action a_i, he bears the cost $C_i(a_i, \theta_i)$, where $\theta_i \in \Theta_i \equiv [\underline{\theta}_i, \overline{\theta}_i]$ is a parameter representing the agent's private information (his type). As a consequence, the cost $C_i(a_i, \theta_i)$ is not observable to anyone except agent i. The cost parameters θ_i are drawn independently from commonly known prior distributions $F_i(\theta_i)$ with positive densities $f_i(\theta_i)$. Each agent has a reservation utility level normalized to zero.

Agents are compensated for the costs they bear by transfer payments x_i made either by the principal or by another agent. All parties are assumed to be risk-neutral. Hence, each agent's utility is the (expected) difference between the transfer payment he receives and his cost. Similarly, the principal seeks to maximize the difference between the benefit $B(a)$ and the sum of the (expected) payments made to agents. We make the following assumptions regarding the structure of production and information:

(A1) $C_i(a_i, \theta_i) = b_i(\theta_i)c_i(a_i)$, where $b_i(\cdot)$ is twice differentiable, strictly positive, strictly increasing, and convex.

(A2) Each agent has the option of not producing at all at zero cost, i.e., there exists $0 \in A_i$ with $c_i(0) = 0$ and $c_i(a_i) > 0$ for all other $a_i \in A_i$.

10. The analysis easily extends to the case of n agents, where authority over contracting and coordination with $n - 1$ agents is delegated to the remaining agent. The case $n = 2$ simplifies the notation considerably, so we adopt this version throughout.

(A3) $\dfrac{b_i'(\theta_i)}{b_i(\theta_i)} \cdot \dfrac{F_i(\theta_i)}{f_i(\theta_i)}$ is increasing in θ_i.

Assumption (A1) postulates that the cost function is multiplicatively separable in the state variable θ_i and the production level. The role of this assumption will become clear subsequently. We do not impose a specific structure for the production sets, except for (A2), which allows each agent to be inactive at zero cost. Finally, assumption (A3) is a variant of the usual requirement that the inverse hazard rate $F_i(\theta_i)/f_i(\theta_i)$ be increasing in θ_i. Basically, (A3) enables us to solve for optimal contracts on the basis of the local conditions for incentive compatibility only. It also ensures that menus of linear contracts are optimal.[11]

It will be useful to recall the nature of optimal mechanisms when the organization faces no contracting constraints. Then the Revelation Principle applies, implying that the principal can without loss of generality restrict himself to centralized revelation mechanisms, where agent i reports his entire private information θ_i to the principal, who subsequently decides transfers $x_i(\theta_1, \theta_2)$ and production assignments $a_i(\theta_1, \theta_2)$. The optimal revelation mechanism solves the following problem:

$$\max_{a(\theta),x(\theta)} E_\theta\left[B(a(\theta)) - \sum_{i=1}^{2} x_i(\theta)\right]$$

subject to: for all $\theta_i \in \Theta_i$, $i \in \{1, 2\}$:

(i) $\theta_i \in \arg\max_{\tilde{\theta}_i} E_{\theta_j}[x_i(\tilde{\theta}_i, \theta_j) - C_i(a_i(\tilde{\theta}_i, \theta_j), \theta_i)]$,

(ii) $E_{\theta_j}[x_i(\theta_i, \theta_j) - C_i(a_i(\theta_i, \theta_j), \theta_i)] \geq 0$.

The first constraint (i) is the standard incentive compatibility condition, requiring that truthtelling be a Bayesian-Nash equilibrium.[12] Constraint (ii) is a participation constraint that reflects the fact that each agent knows his cost prior to contracting.[13]

The solution to the above problem is well known from the literature on adverse selection. For any given production assignments, $a(\theta)$

11. See MMR (1992) for further discussion of this assumption.

12. We use the notation $\theta = (\theta_1, \theta_2)$ and $E_\theta[\cdot] = \int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} [\cdot]\, dF_2(\theta_2)\, dF_1(\theta_1)$. Similarly, $E_{\theta_i}[\cdot] = \int_{\underline{\theta}_i}^{\bar{\theta}_i} [\cdot]\, dF_i(\theta_i)$.

13. Alternatively, agents receive their private information after contracting, and they can costlessly quit after receiving their information.

$\equiv (a_1(\theta_1, \theta_2), a_2(\theta_1, \theta_2))$, the payments to the agents are uniquely determined by the incentive and participation constraints. Furthermore, the agents will earn informational rents because of their private information; i.e., their expected payoffs will generally exceed their reservation utility levels. The principal needs to trade off production efficiency against the informational rents earned by the agents. The optimal balance between these conflicting objectives is summarized by the following result[14]:

LEMMA 0 (MMR, 1995): *Given assumptions* (A1)–(A3), *the production assignments* $a^*(\theta)$ *in the optimal revelation mechanism satisfy*

$$(a_1^*(\theta_1,\theta_2), a_2^*(\theta_1,\theta_2)) \in \arg\max_{(a_1, a_2)} \{B(a_1,a_2) - h_1(\theta_1)c_1(a_1) - h_2(\theta_2)c_2(a_2)\}, \tag{1}$$

where

$$h_i(\theta_i) \equiv b_i(\theta_i)\left(1 + \frac{F_i(\theta_i)}{f_i(\theta_i)} \frac{b_i'(\theta_i)}{b_i(\theta_i)}\right).$$

Expression (1) implies that the principal effectively marks up each agent's unit cost by a factor exactly equal to the modified inverse hazard rate [as defined in assumption (A3)]. The principal calculates optimal production assignments with this measure of ''virtual'' cost, i.e., $h_i(\theta_i)c_i(a_i)$, rather than with the true cost, $b_i(\theta_i)c_i(a_i)$. In expectation, the markup represents the informational rent that agent i earns due to his private information.

3. Limited Contract Contingencies

In this section, we introduce restrictions on the complexity of contracts. We identify the notion of complexity primarily with the number of contingencies in a contract, and secondarily with the number of decisions stipulated in each contingency. These collectively define the *length* of the contract. Our basic notion of complexity is that contracts that involve more contingencies and stipulate more decisions in each contingency are costlier to write and enforce.

To illustrate the notion of limited contract contingencies in the context of a revelation mechanism, note that agent i's contract consists

14. The proof of this result can be found in MMR (1995). All other proofs are contained in the Appendix.

of the two functions $\{a_i(\theta_1, \theta_2), x_i(\theta_1, \theta_2)\}_{(\theta_1,\theta_2)\in\Theta_1\times\Theta_2}$. We interpret the pairs (θ_1, θ_2) as the contingencies of the contract, since the agent's action choice and payment must be specified for all possible combinations of reports. A revelation mechanism therefore involves a continuum of contingencies, and for each contingency the contract specifies two variables.

In the subsequent analysis we impose the exogenous restriction that the number of contract contingencies is finite. This restriction reflects the notion that the cost of preparing and enforcing a contract is increasing in the number of contingencies. While it will not be necessary for our purposes to specify an explicit cost function, we are *de facto* ruling out full revelation mechanisms as prohibitively costly. Our approach is consistent with the notion that in many contracting situations the set of possible contingencies constitutes a large multidimensional set, yet contracts are written on just a few summary variables.[15]

3.1 Centralized Contracting

With limitations on the number of contract contingencies the Revelation Principle no longer applies, and the search for optimal mechanisms becomes substantially more complicated. For instance, it is no longer obvious that both agents should send their reports simultaneously to the principal. Indeed, we shall show below that it is typically better from the principal's perspective for them to report sequentially. Furthermore, mechanisms with sequential reporting could conceivably be dominated by ones with even more complicated message-sending rules, e.g., where agents send reports iteratively. The goal of this paper, though, is to compare the performance of specific organizational arrangements that differ with respect to the allocation of decision rights. In that sense, we are not aiming for a theory of optimal mechanisms when contracts are limited in complexity.

We begin with the natural extension of a revelation mechanism to a setting with finitely many contingencies. Both agents simultaneously send a message to the principal, but each agent is restricted to select messages from a finite set M_i, with $|M_i| \equiv k_i$. The contract stipulates action choices $a_i(m_1, m_2)$ and payments $x_i(m_1, m_2)$ for all possible reports (m_1, m_2). The resulting contract then involves k_1k_2 contingencies,

15. Related ideas have been explored in various strands of the decentralization literature; for instance, Hurwicz (1977, 1987), Mount and Reiter (1974), Jordan (1989) and Baiman (1991).

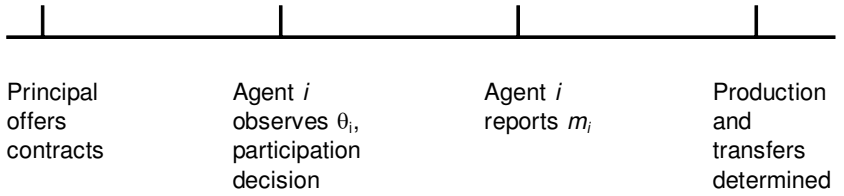

FIGURE 1. TIME LINE 1: Simultaneous Centralized Contracting

with two variables specified for each contingency. The sequence of moves is represented in Figure 1.

We shall refer to these mechanisms as *simultaneous centralized mechanisms*; they stipulate for each agent a reporting (message-sending) rule $\lambda_i : \Theta_i \to M_i$ and a contract $a_i : M_1 \times M_2 \to A_i$, $x_i : M_1 \times M_2 \to \mathbb{R}$. The principal seeks to maximize his expected profit:

$$\max_{\substack{a_i(\cdot),x_i(\cdot)\\ \lambda_i(\cdot)}} E_\theta\left[B(a(\lambda(\theta_1,\theta_2))) - \sum_{i=1}^{2} x_i(\lambda(\theta_1,\theta_2))\right]$$

subject to: for all θ_i, $1 \le i \le 2$,

$$\text{(i)}\quad \lambda_i(\theta_i) \in \arg\max_{m_i \in M_i} E_{\theta_j}[x_i(m_i, \lambda_j(\theta_j)) - b_i(\theta_i)c_i(a_i(m_i, \lambda_j(\theta_j)))],$$

$$\text{(ii)}\quad E_{\theta_j}[x_i(\lambda_i(\theta_i), \lambda_j(\theta_j)) - b_i(\theta_i)c_i(a_i(\lambda_i(\theta_i), \lambda_j(\theta_j)))] \ge 0. \qquad (2)$$

The notation $a(\lambda(\theta_1, \theta_2))$ is shorthand for $(a_1(\lambda_1(\theta_1), \lambda_2(\theta_2)), a_2(\lambda_1(\theta_1), \lambda_2(\theta_2)))$, and λ denotes (λ_1, λ_2).

The two constraints represent the requirement that participating in the mechanism and reporting according to the suggested rules form a Bayes-Nash equilibrium.

The main difference from a revelation mechanism is that agents have to select reports from a finite message set that induces partial pooling for subsets of types. In this sense the complexity constraint restricts the extent to which production assignments and payments can be fine-tuned to variations in the true state of the world. As we show below, the multiplicative separability assumption (A1) ensures that any such mechanism is equivalent to one where the pattern of pooling represents an interval partition of the type space. In other words, if any two types report the same message, then so do all intermediate types. This greatly simplifies our analysis, since we can confine attention to

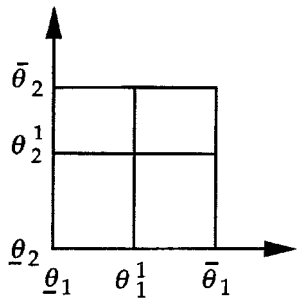

FIGURE 2. SIMULTANEOUS CENTRALIZED CONTRACTING

reporting rules that induce a rectangular partition of the type space, i.e., a grid. Formally, the principal selects for each agent a partition of the type space Θ_i into k_i intervals $(\theta_i^{u-1}, \theta_i^u]$, $u = 1, \ldots, k_i$, with $\theta_i^0 = \underline{\theta}_i$, $\theta_i^{k_i} = \bar{\theta}_i$. All types in the same interval $(\theta_i^{u-1}, \theta_i^u]$ then report the same message $m_i^u \in M_i$, so that the principal can no longer distinguish among them. See Figure 2 for an illustration of the case of four contingencies, with two possible messages per agent. The communication can be interpreted as each agent reporting that costs are either "high" or "low"; more detail cannot be incorporated, owing to the need to limit the complexity of the contract.

In terms of the production assignments, the only difference from a revelation mechanism is that the principal is constrained to select the same assignments for all types in the same interval. Given any assignment, the usual arguments underlying the revenue equivalence theorem (see, for example, Myerson, 1981) apply, and therefore an agent's expected payment has to be equal to the virtual cost of his production assignment. The principal's optimization problem thus reduces to

$$\max_{\substack{\{\theta_1^u,\,\theta_2^v\} \\ \{a_i^{uv}\}}} \sum_{u=1}^{k_1} \sum_{v=1}^{k_2} \int_{\theta_1^{u-1}}^{\theta_1^u} \int_{\theta_2^{v-1}}^{\theta_2^v} \left(B(a_1^{uv}, a_2^{uv}) - \sum_{i=1}^{2} h_i(\theta_i) c_i(a_i^{uv}) \right) dF_2(\theta_2) dF_1(\theta_1), \tag{3}$$

where the variable a_i^{uv} denotes agent i's action choice following reports (m_1^u, m_2^v).

To summarize, the problem of selecting an optimal centralized mechanism with simultaneous reporting and $k_1 k_2$ contract contingencies can be represented as follows: the principal chooses a grid for Θ_1

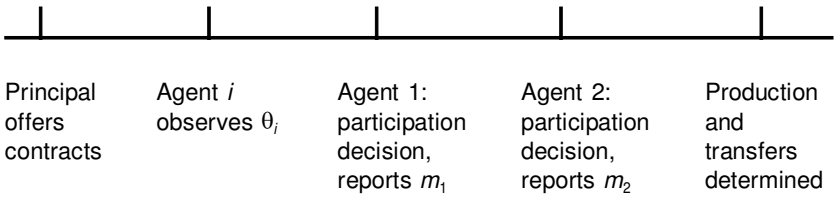

FIGURE 3. TIME LINE 2: Sequential Centralized Contracting

$\times \Theta_2$ consisting of $k_1 k_2$ rectangles. For each rectangle (contingency) an agent's contract specifies a pair (a_i^{uv}, x_i^{uv}). The action choices a_i^{uv} can be chosen optimally, while the corresponding payments, by virtue of the revenue equivalence theorem, are determined uniquely by the incentive and participation constraints.

LEMMA 1: *The value of the centralized contracting problem with simultaneous reporting as stated in* (2) *is equal to the value of the program in* (3).

With limited contract contingencies it is conceivable that the principal can gain from asking the agents to make their reports in sequence rather than simultaneously. The second agent could be instructed to send his message in response to the first agent's report. The exact sequence of events is described in Figure 3. The advantage of sequential reporting is that the second agent can condition his report on the information revealed by the first agent. This additional flexibility will generally induce better coordination of the agents' decisions.

The principal's optimization problem with sequential reporting is similar to that with simultaneous reporting. However, agent 2's reporting rule now takes the form $\lambda_2(\theta_2, \lambda_1(\theta_1))$. Given assumption (A1), the principal can again restrict attention to reporting rules that form an interval partition of the type space of each agent. Sequential reporting allows for the partition of Θ_2 to depend on the report sent by the first agent. Specifically, types in $(\theta_2^{u,v-1}, \theta_2^{uv})$ can send the report m_2^u, following the report m_1^u of the first agent. Figure 4 illustrates the additional flexibility of sequential mechanisms when each agent can send one of two messages.

Contrasted with the increased flexibility is the fact that incentive and participation constraints for agent 2 are strengthened, since agent 2 has the advantage of knowing agent 1's report before responding to

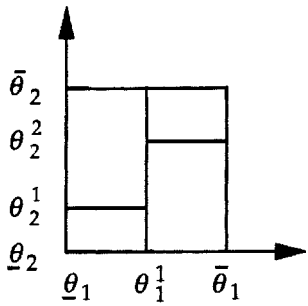

FIGURE 4. SEQUENTIAL CENTRALIZED CONTRACTING

the principal. Specifically, agent 2's incentive compatibility constraint takes the following form:

$$m_2^{uv} \in \arg\max_{m_2 \in M_2} [x_2(m_1^u, m_2) - b_2(\theta_2)c_2(a_2(m_1^u, m_2))]$$

$$\text{for all} \quad 1 \le u \le k_1 \text{ and } \theta_2 \in (\theta_2^{u,v-1}, \theta_2^{uv}).$$

Furthermore, the participation constraint has to hold *ex post*, i.e.,

$$x_2(m_1^u, m_2^{uv}) - b_2(\theta_2)c_2(a_2(m_1^u, m_2^{uv})) \ge 0.$$

It turns out, however, that these seemingly stronger constraints impose no additional cost on the principal. The reason is essentially the same as that described in Mookherjee and Reichelstein (1992); an optimal mechanism in which truthful reporting is a Bayes-Nash equilibrium and the participation constraints apply in an interim sense can be replaced by another mechanism in which truthtelling is a dominant strategy and the participation constraints hold *ex post*. Furthermore the new mechanism yields the same expected payoff to the principal.

We are now in a position to characterize the principal's expected payoff under centralized contracting with sequential reporting.

PROPOSITION 1: *The value of centralized contracting with sequential reporting of k_1 and k_2 messages, respectively, is equal to the value of the following optimization problem:*

$$\max_{\substack{\{\theta_1^u\},\{\theta_2^{uv}\} \\ \{a_i^{uv}\}}} \sum_{u=1}^{k_1} \sum_{v=1}^{k_2} \int_{\theta_1^{u-1}}^{\theta_1^u} \int_{\theta_2^{uv-1}}^{\theta_2^{uv}} \left(B(a_1^{uv}, a_2^{uv}) - \sum_{i=1}^{2} h_i(\theta_i)c_i(a_i^{uv}) \right)$$

$$dF_2(\theta_2)dF_1(\theta_1). \quad (4)$$

Comparison of (3) and (4) immediately shows that the principal is better off with sequential than with simultaneous reporting for any given message sets M_1 and M_2. Problem (4) reduces to (3) if one additionally imposes the constraint that the cutoff values for agent 2's report θ_2^{uv} be independent of Agent 1's report u. As mentioned before, though, there may well exist other centralized contracting arrangements which dominate even (4). Such mechanisms may involve iterative message sending and/or delegation of the action choices once agents have sent their messages. The centralized contracting scenario we consider is nonetheless of considerable interest, since it is the direct analogue of a revelation mechanism in a setting with limited contract contingencies.

3.2 Delegated Contracting

Consider now a contracting arrangement wherein the principal only contracts with agent 1, who in turn is authorized to subcontract with agent 2. This arrangement amounts to a three-tier hierarchy in which agent 1 acts as an intermediate principal (a ''manager'' or ''prime contractor''). As in many contexts of procurement contracting, agent 1 (the prime contractor) pays agent 2 (the subcontractor) out of her own pocket, and this payment cannot be monitored by the principal. However, as explained above, we assume that the principal does monitor the allocation of production assignments between the two agents.

When contracting is costly and necessarily incomplete, a three-tier hierarchy possesses one potential advantage. Agent 1 can design the subcontract for agent 2 on the basis of full information about her own cost. This may result in improved decision making when compared to a centralized arrangement in which decisions can only be based on agent 1's limited communication. On the other hand, a potential drawback of delegated contracting is that the principal may experience a *control* loss, owing to the monopsony power granted to agent 1. This problem has been identified in various models of hierarchical contracting, including McAfee and McMillan (1995) and Qian (1994). Our earlier work (MMR, 1995) has shown that in the absence of contracting constraints, the principal can alleviate the control loss completely by constructing a suitable subsidy for outsourcing to agent 2. To calibrate the subsidy correctly, however, the principal has to know agent 1's true cost state θ_1, as this determines the magnitude of the monopsony distortion. To elicit this information, however, the principal has to offer agent 1 a contract with a continuum of contingencies, corresponding to different possible true values of θ_1. This is of course not feasible in the present setting. Hence the control loss cannot be eliminated with a limit on the number of contract contingencies.

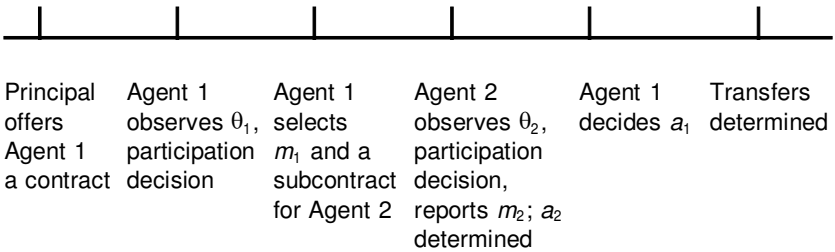

FIGURE 5. TIME LINE 3: Delegated Contracting

The sequence of events under delegation is described in Figure 5. In particular, we assume that agent 1 can commit to the prime contract before entering into a subcontract with agent 2.[16]

The principal designs a contract for agent 1 that stipulates the payment x_1 as a function of the benefit level B delivered, the manager's own contribution a_1, and m_1, a message that agent 1 sends to the principal at the time of contracting. The message m_1 is selected from a finite set M_1, thus allowing the manager to self-select into different incentive contracts depending on his private information. Following the selection of a contract for himself, the manager selects a pair of functions $\{x_2(m_2), a_2(m_2)\}_{m_2 \in M_2}$ specifying the payment that the manager will make to agent 2 and the associated action choice.

The number of contingencies in the subcontract clearly equals $k_2 = |M_2|$, and for each contingency the subcontract specifies two variables, just as in the centralized arrangement. The contract for agent 1 is, however, different in that it must be conditioned on action choices made by her. Hence limits must be imposed on the range of actions that agent 1 is permitted to select from. For each message m_1, the principal can specify a control set $S(m_1)$ such that

$$S(m_1) \subset A_1 \times \{B(a_1, a_2) \mid a_1 \in A_1, a_2 \in A_2\}. \tag{5}$$

The interpretation of this control set is that, having selected m_1, agent 1 is contractually obligated to deliver a combination of a_1 and B that belongs to the finite set $S(m_1)$. The number of contingencies in the prime

16. MMR (1995) demonstrated that this sequence of contracting is the most beneficial to delegation.

contract is then given by $\sum_{m_1 \in M_1} |S(m_1)|$.[17] In particular, if $|M_1| = k_1$, and $|S(m_1)| = k_2$ for all m_1, then the prime contract involves $k_1 k_2$ contingencies. In this manner, restricting the number of contingencies in the delegation contract limits the range of production assignments as well as reports that agent 1 can select from. In contrast, the setting of limited communication studied in MMR (1995) did not impose any limits on the range of production assignments. Hence delegation is more constrained when it is contractual complexity rather than communication capacity which is restricted.

Given the prime contract, agent 1's type θ_1 and his message m_1, we denote agent 1's payoff by

$$\pi_1(a_1, a_2, x_2 \mid x_1(\cdot), m_1, \theta_1) \equiv x_1(B(a_1, a_2), a_1, m_1) - x_2 - b_1(\theta_1)c_1(a_1).$$

The manager's subsequent problem of designing a subcontract is to select for agent 2 a reporting rule $\lambda_2(\theta_2) : \Theta_2 \to M_2$ and functions $a_1(m_2)$, $a_2(m_2)$, $x_2(m_2)$ to maximize

$$E_{\theta_2}[\pi_1(a_1(\lambda_2(\theta_2)), a_2(\lambda_2(\theta_2)), x_2(\lambda_2(\theta_2)) \mid x_1(\cdot), m_1, \theta_1)] \tag{6}$$

subject to the constraints that for all $\theta_2 \in \Theta_2$,

$$\lambda_2(\theta_2) \in \arg\max_{m_2 \in M_2} [x_2(m_2) - b_2(\theta_2)c_2(a_2(m_2))],$$

$$x_2(\lambda_2(\theta_2)) - b_2(\theta_2)c_2(a_2(\lambda_2(\theta_2))) \geq 0,$$

$$(a_1(m_2), B(a_1(m_2), a_2(m_2))) \in S(m_1) \qquad \text{for all} \quad m_2 \in M_2.$$

The choice of subcontract will generally depend on the prime contract and agent 1's type θ_1. In the terminology of Maskin and Tirole (1990), the subcontracting problem involves an informed principal problem *with private values*: i.e., agent 1's valuation of the output delivered by agent 2 depends on θ_1, regarding which agent 1 is privately informed. Nevertheless, this variable exercises no direct effect on agent 2's utility. Due to the assumed risk neutrality, however, the informed-principal problem has essentially no effect, and we can solve for the optimal subcontract as if agent 1's type were common knowledge between the two agents.[18]

17. We are implicitly assuming here that the contract states that agent 1 will be paid nothing, unless one of the production combinations in $S(m_1)$ is delivered, and this exclusion clause is costless to write into the contract, or alternatively is equivalent to the cost of writing a single contingency.

18. As Maskin and Tirole demonstrate in their paper, this is generally true when both parties have utility that is linear in money. The only advantage that agent 1 could conceivably extract from his private information is to delay revelation of his own type θ until after agent 2 responds to the offered contract, thereby imposing risk on agent 2 associated with the realization of θ_1. With risk neutrality this serves no purpose at all.

We denote the principal's maximum payoff under centralized contracting with sequential reporting and $k \equiv k_1k_2$ contingencies for both agents by $\pi^c(k_1k_2, k_1k_2)$. Formally, $\pi^c(k_1k_2, k_1k_2)$ is the optimal value of the objective function in (4). The principal's maximum payoff under delegated contracting is denoted by $\pi^d(\tilde{k}, k^*)$ when the prime contract and subcontract involve $\tilde{k}$ and k^* contingencies, respectively. The main result of this paper is the following.

PROPOSITION 2: *Centralized contracting with sequential reporting is dominated by delegated contracting in the sense that* $\pi^c(k_1k_2, k_1k_2) \leq \pi^d(k_1k_2, k_2)$ *for all* k_1 *and* k_2.

Note that in comparison with the corresponding centralized contract, delegation involves fewer contingencies in the contract for agent 2, while the number of contingencies in agent 1's contract remains the same. The message sets of all agents have the same size: hence the result here implies also the superiority of delegation when only communication is limited, as in MMR (1992). Moreover, the subcontract specifies the same variables per contingency, i.e., the production assignment and the payment for agent 2. For agent 1, a contingency is defined by the realization of the three variables B, a_1, m_1. For each contingency the prime contract stipulates just one variable, i.e., x_1, rather than the two variables (x_1, a_1) under centralization. Hence the delegation mechanism is less complex than a centralized mechanism, judging not only by the total number of contingencies, but also by the number of decisions the principal takes in each contingency.

The proof of Proposition 2 is constructive. Given a centralized contract, the principal can design the prime contract under delegation so as to leave agent 1's message space M_1 and the domain of possible production decisions $\{a_1^{u^v}, a_2^{u^v}\}$ unchanged. Given that agent 1 sends message m_1^u, his discretion over production decisions is restricted exactly to the set of possible production assignments resulting under centralization following the message m_1^u. In other words, the control set $S(m_1^u)$ is the set of k_2 possible production assignments $\{a_1^{u^v}, a_2^{u^v}\}_{v=1}^{k_2}$ in the centralized contract. Despite this restriction, though, agent 1 now decides on the production assignments with full knowledge of his own type θ_1, while under centralization the principal only knows that agent 1's cost belongs to the interval $(\theta_1^{u-1}, \theta_1^u]$. This feature entails a flexibility gain for delegation.[19]

19. Note that this advantage would continue to exist even if the centralized regime were permitted to use iterative reporting schemes. While the latter increase the flexibility of the centralized arrangement by narrowing down the uncertainty faced by the principal concerning agent 1's type, they cannot eliminate it entirely, owing to the finite restriction on message spaces.

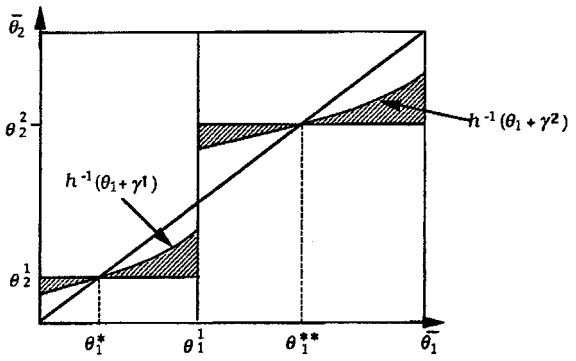

FIGURE 6. SUPERIORITY OF DELEGATION OVER CENTRALIZATION

On the other hand, recall that the control loss problem inherent in delegation is that agent 1 has a tendency to use his monopsony power to bias production in his favor when allocating tasks between agent 2 and himself. The principal can counteract this distortion by subsidizing procurement from agent 2. The limitation on the number of contract contingencies however restricts the set of reports that agent 1 can send regarding his own type, thus preventing the principal from calibrating the outsourcing subsidy correctly. Nevertheless, we find that the control loss can be ameliorated sufficiently so as to induce agent 1 to make uniformly ''better'' decisions than would have resulted under centralized contracts with the same number of contingencies.

To further illustrate the reasoning underlying Proposition 2, consider an example in which there are two competing suppliers with cost functions $C_i(a_i, \theta_i) = \theta_i a_i$, respectively. Suppose both θ_i's are distributed according to the same density $f(\theta_i)$, resulting in common virtual cost functions $h(\theta_i)$. The benefit function $B(\cdot, \cdot)$ takes the form $B(a_1, a_2) = \max\{a_1, a_2\}$, and the benefit level is exogenously fixed at $\bar{B} = 1$. The optimal revelation mechanism then calls for agent 1 to deliver $a_1 = 1$ if and only if $h(\theta_1) \leq h(\theta_2)$, which is equivalent to $\theta_1 \leq \theta_2$. Otherwise agent 2 is asked to deliver $a_2 = 1$, so the low-cost producer is given the entire contract.[20] The optimal production assignments are thus described by the diagonal in Figure 6: this necessitates

20. It can be shown that in this setting the optimal revelation mechanism can be implemented by a second price auction.

each agent reporting his entire private information, i.e., the realization of θ_i. In turn this causes the corresponding contract to contain a continuum of contingencies.

Consider now a scenario in which there are four contingencies in the centralized contract, with two messages from each agent. It is then no longer possible to implement second-best production assignments, as the principal cannot identify which producer has the lower cost in all states of the world. We know from Proposition 1 that the optimal centralized contract with sequential reporting will take the form of a rectangular partition of the cost parameter space, rather than the diagonal partition. Heuristically, it will involve choosing the rectangular partition that is the closest approximation to the diagonal partition. In particular, if agent 1 has cost $\theta_1 \leq \theta_1^1$, he reports m_1^1. Hearing this, agent 2 reports m_2^1 if $\theta_2 \leq \theta_2^1$ and m_2^2 otherwise. In the former case agent 2 produces the entire quantity, and in the latter case agent 1 does. If $\theta_1 > \theta_1^1$, agent 1 reports m_1^2. Subsequently agent 2 reports m_2^1 if $\theta_2 \leq \theta_2^2$ and m_2^2 otherwise. The values of θ_2^1 and θ_2^2 are also the prices at which agent 2 is offered the contract depending on agent 1's message. The cutoff levels θ_2^1 and θ_2^2 are such that for some type (θ_1^* or θ_1^{**}) of agent 1 in the corresponding interval, production decisions are exactly second-best (see Figure 6).

In delegated contracting, the principal can use the following mechanism to achieve higher profits. Agent 1 is asked to use the same reporting rule as in the two-tier mechanism. Following message m_1^u, he is offered an incentive scheme of the form $x_1^u = \beta^u - a_1\gamma^u$, where β^u is a fixed payment, and γ^u is a "tax" parameter chosen as follows: $\gamma^1 = h(\theta_1^*) - \theta_1^*$ and $\gamma^2 = h(\theta_1^{**}) - \theta_1^{**}$. Consequently, agent 1 will make a take-it-or-leave-it offer to agent 2 at the price $h^{-1}(\theta_1 + \gamma^1)$ if $\theta_1 \leq \theta_1^1$, or, at the price $h^{-1}(\theta_1 + \gamma^2)$ if $\theta_1 > \theta_1^1$. Hence, the unit tax γ^i mitigates the control loss, since in the absence of monitoring agent 1 would have chosen the price $h^{-1}(\theta_1)$. By construction, the coefficients γ^1 and γ^2 are such that the types θ_1^* and θ_1^{**} continue to implement the second-best decisions. Figure 6 shows that the production assignments corresponding to the $h^{-1}(\theta_1 + \gamma^i)$ rule are *uniformly* closer to the diagonal than the production assignments in the optimal centralized mechanism as represented by the flat lines corresponding to the constant prices θ_1^1 and θ_2^2. As a consequence, the production decisions are uniformly better from the principal's standpoint.

When communication rather than the number of contract contingencies is limited, as in our previous paper MMR (1992), the only constraint pertains to the size of each agent's message set; production decisions are not restricted at all. Agent 1 can then select arbitrary

production assignments under delegation, based on his own private information, besides reports from agent 2. The flexibility advantage of delegation is then further boosted: a continuum of different assignments can be selected by agent 1, corresponding to different values of θ_1, given any report sent by agent 2. In the preceding example, however, this additional flexibility was not valuable, since the aggregate benefit level desired by the principal was fixed exogenously. More generally, if the principal's aggregate benefit were a continuous function of team output, the added flexibility in production assignments would enable achievement of a higher expected payoff.

The contrast between the context of limited communication and limited contract contingencies can be illustrated more sharply in a single-agent setting. Suppose the only restriction is on the size of the agent's message space; then delegation of decisions concerning production to the agent would generally generate superior performance to centralization. This is because delegation permits a continuum of possible production decisions that are fine-tuned to the agent's private information, unlike centralization, where the principal decides on production based on reports from the agent. On the other hand, when the only constraint is on the number of contract contingencies, centralized and delegated decision making are equivalent. The reason is that delegation also must be characterized by a finite range of possible production levels that can be selected by the agent, owing to the need to limit the number of contingencies in the delegation contract.

In the example used above to illustrate the reasoning of Proposition 2, delegation achieved a strictly higher expected payoff for the principal than did centralization. The following result provides sufficient conditions for this to hold in more general settings. In stating the result, we shall need the following definitions. First, say that *communication with agent i is valuable under centralization* if the principal's payoff in (4) increases as one moves from $k_i = 1$ to some $k_i > 1$ (holding k_j, $j \neq i$, fixed). Second, given a vector of unit prices $p = (p_1, p_2)$ for the two inputs, let $a(p)$ denote the vector of input levels (a_1, a_2) that maximizes $B(a) - p_1c_1(a_1) - p_2c_2(a_2)$. We define the production structure to be *nonseparable* if for any pair of input prices p, p' both lying in $[h_1(\underline{\theta}_1), h_1(\overline{\theta}_1)] \times [h_2(\underline{\theta}_2), h_2(\overline{\theta}_2)]$, and any corresponding pair of optimal production assignments $a(p), a(p')$, if $a_2(p) \neq a_2(p')$, then also $a_1(p) \neq a_1(p')$. In words, when the production contribution procured from one agent changes, so must the contribution procured from the other agent. This condition thus requireness jointness in the production of the

two agents, necessitating coordination of their respective production assignments.[21]

PROPOSITION 3: *Suppose that communication with agent* 2 *is valuable under centralization, and the production structure is nonseparable. Then the dominance of delegated contracting over centralization in Proposition* 2 *is strict, i.e.,* $\pi^c(k_1k_2, k_1k_2) < \pi^d(k_1k_2, k_2)$ *for all* k_1, k_2.

To conclude this section, we illustrate the importance of the assumption that the principal can monitor the production assignments selected by agent 1. If the principal were to consider only the aggregate benefit level delivered, but not agent 1's contribution a_1, the control loss might outweigh the flexibility gain inherent in delegation. To demonstrate this point, consider again the above procurement example where the two agents are *ex ante* identical and the desired benefit level is fixed exogenously at $\overline{B} = 1$. The optimal delegation mechanism will then be of the following simple form: the principal offers agent 1 a fixed payment in the amount of

$$\overline{x}_1 = h^{-1}(\overline{\theta}_1)F(h^{-1}(\overline{\theta}_1)) + \overline{\theta}_1[1 - F(h^{-1}(\overline{\theta}_1))]. \tag{7}$$

The amount $\overline{x}_1$ is calculated to exactly compensate the highest cost type $(\overline{\theta}_1)$ of agent 1 for delivery cost incurred.

This prime contract induces agent 1 to make agent 2 a take-it-or-leave-it offer to produce the entire assignment at the price $h^{-1}(\theta_1)$. Hence agent 2 ends up as the exclusive producer if and only if $\theta_2 \leq h^{-1}(\theta_1)$. The flexibility of the three-tier mechanism is embodied in this decision rule, since the allocation of production is decided on the basis of agent 1's exact information about his own cost θ_1. This flexibility gain, however, is not necessarily desirable, since the resulting decision rule is distorted relative to the second-best rule.

The following sequential centralized mechanism performs better than the above delegation mechanism. Agent 1 is asked to report m_1^1 if $\theta_1 \leq \theta_1^1$. The principal awards the entire contract to agent 2 if $\theta_2 < h^{-1}(\theta_1^1) \equiv \theta_2^1$, and to agent 1 otherwise. If agent 1 reports m_1^2 (i.e., $\theta_1 > \theta_1^1$), the contract is awarded to agent 2 if and only if $\theta_2 \leq h^{-1}(\overline{\theta}_1)$

21. Note that if the optimal production assignments of the two agents were completely separable, e.g., if the principal's benefit function B were additively separable in a_1 and a_2, then delegation could not lead to any flexibility gain at all: better information about the cost conditions in the production of one agent would not generate any improvements in the contract designed for the other agent. Hence *some* degree of nonseparability is necessary for delegation to achieve more flexible production assignments than centralization.

(i.e., agent 2 reports m_2^1). The principal's expected cost from this centralized mechanism equals

$$K(\theta_1^1)F(\theta_1^1) + K(\overline{\theta}_1)\,[1 - F(\theta_1^1)], \tag{8}$$

where $K(\theta_1) \equiv h^{-1}(\theta_1)F(h^{-1}(\theta_1)) + \theta_1[1 - F(h^{-1}(\theta_1))]$. The expected cost in (8) is clearly less than in (7), i.e., with $\theta_1^1 = h^{-1}(\theta^-)$ and $\overline{x}_1 = K(\overline{\theta}_1)$; therefore the principal is better off under centralization. By adopting a centralized mechanism, the principal loses some production efficiency associated with delegation. At the same time, though, types $\theta_1 < \theta_1^1$ earn lower rents in the three-tier hierarchy, and the resulting expected cost is lower under centralization. Put differently, the advantage of flexibility under delegation is appropriated by agent 1 rather than by the principal.

4. Concluding Remarks

Our analysis has shown that with limited contract contingencies a principal will generally benefit from delegating authority to coordinate production and to contract with other agents. By designing a suitable prime contract the principal can align her own preferences with those of the manager sufficiently so as to take advantage of the flexibility inherent in delegation. These results are consistent with the widespread prevalence of managerial hierarchies and with the practice of subcontracting. At the same time, our results point to several features that appear essential in order for a subcontracting arrangement to indeed be superior.

Our model ignored a number of factors that may favor centralization over delegation. These include the possibility of collusion between agents; the presence of limited-liability constraints on agents, which prevent intermediate contractors from bearing too much risk; and the inability of the principal to ensure the appropriate sequencing of contracts under delegation.[22] In a setting where contract complexity restrictions are absent and the Revelation Principle applies, the importance of these factors for the performance of delegation arrangements has been established in MMR (1995). When there are restrictions on contract complexity and some of the above conditions are not satisfied, the organization designer is likely to face the following trade-off: delegation is beneficial in that it enables production decisions to be more flexible,

22. Collusion between agents or inappropriate sequencing may involve the agents entering into a side contract *before* agent 1 responds to the contract offered by the principal. This expands the extent of private information of agent 1, and therefore also the rents that the latter can capture. In a similar vein, *ex post* limited liability constraints on agent 1 would also enable her to capture larger rents. See MMR (1995) for further details.

while centralization allows better control. Further research into the nature of this trade-off is necessary.

We restricted attention to a setting with only two production agents, and thereby to a comparison of a two-tier hierarchy (centralization) with one involving three tiers and a single branch (delegation). With more than two agents, the designer can select from a larger set of multitier hierarchies, varying both in the number of vertical layers and in the number of branches. When a hierarchy contains multiple branches, the need to coordinate their decisions also arises. Mookherjee and Reichelstein (1995, 1997) examine such issues in a setting involving no restrictions on contract complexity or communication. They show that any multitier hierarchy consistent with the technology in a suitable sense continues to achieve optimal outcomes, equivalent to those under centralization. Hence a large variety of structures attain equivalent performance. Once contracts are restricted in complexity, a nontrivial choice amongst different hierarchial structures arises. We know from the results of this paper that centralization is dominated by a specific three-tier hierarchy with a single branch, but comparison of the latter with other structures remains to be explored.

The optimal amount of decentralization in large firms continues to be the subject of much debate in the management literature. Our concern for contract complexity is but one of many factors that are likely to affect the design of organizational subunits with decision-making responsibility. In particular, our model lacks an explicit recognition of the time managers have available for the purpose of processing information, coordinating decisions, and actually carrying out the decisions.

Recent papers by Mount and Reiter (1995), Radner (1993), Radner and van Zandt (1992, 1995), Reiter (1996), and Bolton and Dewatripont (1994) propose a variety of models that all take into consideration the amount of time it takes an organization to arrive at decisions. While these models are explicit in the amount of information an agent can communicate and process in a given amount of time, they have ignored incentive issues thus far. Ultimately, however, a satisfactory theory of organization design must address both incentive considerations and limited information-processing capabilities.

Appendix

Proof of Lemma 1: We prove first that there is no loss of generality in restricting attention to interval partitions. Given any message $m_i^u \in M_i$, define

$\Theta_i^u \equiv \{\theta_i \mid \lambda_i(\theta_i) = m_i^u\}$,

$\underline{\theta}_i^u \equiv \inf \Theta_i^u$, $\overline{\theta}_i^u \equiv \sup \Theta_i^u$, and

$M_i^u \equiv \{m_i \in M_i \mid m_i = \lambda_i(\theta_i), \underline{\theta}_i^u \le \theta_i \le \overline{\theta}_i^u\}$.

STEP 1: *All types in* $[\underline{\theta}_i^u, \overline{\theta}_i^u]$ *are indifferent between all messages in* M_i^u.

Proof: Take any two types, θ_i, θ_i' in $[\underline{\theta}_i^u, \overline{\theta}_i^u]$ with $\lambda_i(\theta_i) = \lambda_i(\theta_i') = m_i$. Suppose there is $\theta_i'' \in (\theta_i, \theta_i')$ with $\lambda_i(\theta_i'') = m_i' \neq m_i$. We show that types θ_i, θ_i'', and θ_i' are all indifferent between the messages m_i and m_i'.

Let $X \equiv E_{\theta_j}[x_i(m_i', \lambda_j(\theta_j))] - E_{\theta_j}[x_i(m_i, \lambda_j(\theta_j))]$ and $Y \equiv E_{\theta_j}[c_i(a_i(m_i', \lambda_j(\theta_j))) - c_i(a_i(m_i, \lambda_j(\theta_j)))]$. Then the incentive constraints imply

$$b_i(\theta_i)Y \ge X, \qquad b_i(\theta_i'')Y \le X, \qquad b_i(\theta_i')Y \ge X.$$

Suppose $Y > 0$. Then $b_i(\theta_i'')Y > b_i(\theta_i)Y$, as b_i is strictly increasing, which contradicts the first two inequalities. If $Y < 0$, we similarly get a contradiction of the last two inequalities. Hence $Y = 0$, upon which the inequalities imply $X = 0$, and all three hold as equalities. In other words, all three types θ_i, θ_i'', and θ_i' are indifferent between the messages m_i and m_i', establishing our claim. □

Delete "unused" messages from M_i to form $\hat{M}_i \equiv \{m_i \in M_i \mid m_i = \lambda_i(\theta_i)$ for some $\theta_i \in \Theta_i\}$, and reorder elements of $\hat{M}_i$ so that $\underline{\theta}_i^u \le \underline{\theta}_i^{u+1}$ for any $m_i^u \in \hat{M}_i$. Let $\hat{k}_i \equiv |\hat{M}_i|$; obviously $\hat{k}_i \le k_i$.

Now construct a new (interval) partition $\{\hat{\theta}_i^u\}_{u=1}^{\hat{k}_i}$ such that $F_i(\hat{\theta}_i^u) - F_i(\hat{\theta}_i^{u-1}) = \text{Prob}[\{\theta_i \in \Theta_i \mid \lambda_i(\theta_i) = m_i^u\}] \equiv \int_{\Theta_i^u} dF_i$. Choose the reporting rule $\hat{\lambda}_i(\theta_i) = m_i^u \in \hat{M}_i$ if $\theta_i \in (\hat{\theta}_i^{u-1}, \hat{\theta}_i^u]$. Note that this implies

$$\text{Prob}[\{\theta_i \in \Theta_i \mid \lambda_i(\theta_i) = m_i^u\}] = \text{Prob}[\{\theta_i \in \Theta_i \mid \hat{\lambda}_i(\theta_i) = m_i^u\}] \equiv \phi_i^u, \tag{i}$$

i.e., the probability that message m_i^u is reported is unchanged.

STEP 2: *If the reporting rules* $\lambda_1(\theta_1)$ *and* $\lambda_2(\theta_2)$ *are replaced by* $\hat{\lambda}_1(\theta_1)$, $\hat{\lambda}_2(\theta_2)$ *respectively, then incentive and participation constraints continue to be satisfied, while the principal's expected welfare is unaffected.*

Proof: Suppose we alter agent i's reporting rule from $\lambda_i(\theta_i)$ to $\hat{\lambda}_i(\theta_i)$. Agent j's expected payoff when reporting message $m_j \in M_j$ is

$$\sum_u \phi_i^u[x_j(m_i^u, m_j) - C_j(a_j(m_i^u, m_j), \theta_j)],$$

and (i) implies that agent j's payoffs are unaffected by the switch in i's reporting rule. Similarly, if agent j reports m_j, the principal's expected welfare is

$$\sum_u \phi_i^u \left(B(a(m_i^u, m_j)) - \sum_{l=1}^{2} x_l(m_i^u, m_j) \right),$$

which is similarly unaltered.

It remains to show that agent i's incentive and participation constraints are satisfied under the new reporting rule $\hat{\lambda}_i(\theta_i)$. Consider any type θ_i in $(\hat{\theta}_i^{u-1}, \hat{\theta}_i^u)$. By construction, $\hat{\theta}_i^{u-1} \geq \underline{\theta}_i^u$. Furthermore, $\hat{\theta}_i^u \geq \max\{\overline{\theta}_i^{u-1}, \overline{\theta}_i^u\}$. (If this is not the case, then $\sum_{l=1}^{u} \text{Prob}[\{\theta_i \in \Theta_i \mid \lambda_i(\theta_i) = m_i^l\}] = F_i(\hat{\theta}_i^u) > \max\{F_i(\overline{\theta}_i^{u-1}), F_i(\overline{\theta}_i^u)\} \geq \sum_{l=1}^{u} \text{Prob}[\{\theta_i \in \Theta_i \mid \lambda_i(\theta_i) = m_i^l\}]$, which contradicts (i).) Therefore, we have $\max\{\overline{\theta}_i^{u-1}, \overline{\theta}_i^u\} > \theta_i > \underline{\theta}_i^u$. Suppose $\overline{\theta}_i^u \geq \theta_i > \underline{\theta}_i^u$. Let $\lambda_i(\theta_i) = m_i^v$, i.e., θ_i chooses the message m_i^v in the original reporting rule. Since there exists $\tilde{\theta}_i \in [\underline{\theta}_i^u, \overline{\theta}_i^u]$ with $\lambda_i(\tilde{\theta}_i) = m_i^u$, application of Step 1 yields the result that type θ_i is indifferent between messages m_i^u and m_i^v.

Finally, suppose $\theta_i > \overline{\theta}_i^u > \underline{\theta}_i^u$ and $\lambda_i(\theta_i) = m_i^v$. Then by the above argument $\overline{\theta}_i^{u-1} > \theta_i > \underline{\theta}_i^u$, and by the chosen reordering of messages, $\underline{\theta}_i^u > \underline{\theta}_i^{u-1}$. Therefore, by Step 1, θ_i is indifferent between m_i^v and m_i^{u-1}. Since by hypothesis $\max\{\overline{\theta}_i^{u-1}, \overline{\theta}_i^u\} = \overline{\theta}_i^{u-1}$, we have $\overline{\theta}_i^{u-1} \geq \overline{\theta}_1^u$ and $\overline{\theta}_i^u > \underline{\theta}_i^u > \underline{\theta}_i^{u-1}$; it follows that message m_i^u is used by some types between $\overline{\theta}_i^{u-1}$ and $\underline{\theta}_i^{u-1}$. So type θ_i is indifferent between m_i^{u-1} and m_i^u, and therefore between m_i^u and m_i^v.

To complete the proof of the lemma, we note that with interval partitions, the principal's objective function reduces to (3) if one incorporates the local incentive constraints. Hence, it remains to show that the solution to (3) is globally incentive-compatible. For this, it suffices to show that the solution to (3) satisfies the monotonicity conditions $a_1^{u-1,v} \geq a_1^{uv}$ and $a_2^{u,v-1} \geq a_2^{u,v}$. Given (A1), the optimization problem in (9) requires the choice of $\{a_1^{uv}, a_2^{uv}\}$ that maximize

$$B(a_1, a_2) - h_1^u c_1(a_1^{uv}) - h_2^v c_2(a_2^{uv}),$$

where $h_1^u \equiv \int_{\theta_1^{u-1}}^{\theta_1^u} h_1(\theta_1)\, dF_1(\theta_1) / [F_1(\theta_1^u) - F_1(\theta_1^{u-1})]$ (with h_2^v defined in

a symmetric fashion). Since $h_i(\theta_i)$ is increasing in θ_i, it follows that $h_1^u \geq h_1^{u-1}$ and $h_2^v \geq h_2^{v-1}$. Hence, $a_1^{u-1,v} \geq a_1^{uv}$, and $a_2^{u,v-1} \geq a_2^{uv}$, completing the proof. □

Proof of Proposition 1: The following optimization program corresponds to the centralized contracting arrangement with sequential reporting and $k_1 k_2$ contingencies per contract:

$$\max_{\substack{a_i(\cdot), x_i(\cdot) \\ \lambda_1(\cdot), \lambda_2(\lambda_1, \cdot)}} E_\theta \left[B(a(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2))) - \sum_{i=1}^{2} x_i(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2)) \right]$$

subject to

$$\lambda_1(\theta_1) \in \arg\max_{m_1 \in M_1} E_{\theta_2}[x_1(m_1, \lambda_2(m_1, \theta_2)) - C_1(a_1(m_1, \lambda_2(m_1, \theta_2)), \theta_1)]$$

$$\text{for all} \quad \theta_1 \in \Theta_1, \qquad \text{(i)}$$

$$\lambda_2(m_1, \theta_2) \in \arg\max_{m_2 \in M_2} \{x_2(m_1, m_2) - C_2(a_2(m_1, m_2), \theta_2)\}$$

$$\text{for all} \quad m_1 \in M_1, \quad \theta_2 \in \Theta_2, \qquad \text{(ii)}$$

$$E_{\theta_2}[x_1(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2)) - C_1(a_1(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2)), \theta_1)] \geq 0$$

$$\text{for all} \quad \theta_1 \in \Theta_1, \qquad \text{(iii)}$$

$$x_2(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2)) - C_2(a_2(\lambda_1(\theta_1), \lambda_2(\lambda_1(\theta_1), \theta_2)), \theta_2) \geq 0$$

$$\text{for all} \quad \theta_1 \in \Theta_1, \quad \theta_2 \in \Theta_2. \qquad \text{(iv)}$$

Requirement (ii) in the above program says that agent 2 should have an incentive to report message $m_2 = \lambda_2(m_1, \theta_2)$ if his type is θ_2 and agent 1 reported m_1. As in the case of simultaneous mechanisms, it can be shown that because of multiplicative separability assumption (A1) one can confine attention to interval partitions such that $(\theta_1^{u-1}, \theta_1^u] = \{\theta_1 \mid \lambda_1(\theta_1) = m_1^u\}$, $1 \leq u \leq k_1$. Similarly, the interval Θ_2 is partitioned into k_2 intervals, depending on the message m_1 of agent 1:

$$(\theta_2^{u,v-1}, \theta_2^{u,v}] = \{\theta_2 \mid \lambda_2(m_1^u, \theta_2) = m_2^v\}.$$

From here on, the steps in the proof are analogous to those in the proof of Lemma 1. Global incentive compatibility of the solution follows from the fact that the expected production costs are decreasing in each agent's type. □

Proof of Proposition 2: Let the optimal centralized contract involve the partition $\{\theta_1^u, \theta_2^{uv}\}$ of the message sets of the two agents, with associated production assignments $\{a^{uv}\}$. Thus, if agent 1's type is $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$ and agent 2's type is $\theta_2 \in (\theta_2^{u,v-1}, \theta_2^{uv}]$, then the former sends the message m_1^u, the latter sends m_2^v, and the production assignments are (a_1^{uv}, a_2^{uv}). Conditional on $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$, the principal's payoff is

$$\pi^c(k_1 k_2, k_1 k_2 \mid \theta_1) = \sum_{v=1}^{k_2} [B(a^{uv}) - h_1(\theta_1) c_1(a_1^{uv}) - h_2^{uv} c_2(a_2^{uv})] \Delta F_2^{uv}, \qquad \text{(i)}$$

where $\Delta F_2^{uv} \equiv F_2(\theta_2^{uv}) - F_2(\theta_2^{u,v-1})$ and h_2^{uv} denotes the expected virtual cost of agent 2 conditional on the event that $\theta_2 \in (\theta_2^{u,v-1}, \theta_2^{uv}]$, i.e.,

$$h_2^{uv} \equiv \int_{\theta_2^{uv-1}}^{\theta_2^{uv}} h_2(\theta_2)\, dF_2(\theta_2)/\Delta F_2^{uv}.$$

Now consider the following prime contract under delegation. The principal chooses the same partition $\{\theta_1^u\}$ for agent 1 as under centralization, and the control sets are defined as follows:

$$S(m_1^u) = \{(a_1, B) \mid a_1 = a_1^u, B = B(a_1^{uv}, a_2^{uv}) \qquad \text{for some} \quad 1 \le v \le k_2\}.$$

Thus agent 1 can choose any combination of production assignments that the principal might have chosen under centralization. Furthermore, the prime contract for agent 1 is chosen as follows:

$$x_1(B, a_1, m_1^u) = \beta_1^u + B - \gamma_1^u c_1(a_1),$$

where γ_1^u will be specified below, and the corresponding fixed payment β_1^u ensures that marginal types earn the required informational rents. Suppose that type $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$ does indeed report m_1^u (we verify this claim at the end of the proof). Since $c_2(a_2^{uv}) \ge c_2(a_2^{u,v+1})$, it is optimal for type $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$ of agent 1 to select an interval partition denoted by $\{\theta_2^v(\theta_1)\}$ and the production assignment a^{uv} when agent 2 reports m_2^v, i.e., $\theta_2 \in (\theta_2^{u,v-1}(\theta_1), \theta_2^{uv}(\theta_1)]$. Agent 1's expected payoff then becomes

$$\sum_{v=1}^{k_2} \{B(a^{uv}) - [b_1(\theta_1) + \gamma_1^u]c_1(a_1^{uv}) - h_2^{uv(\theta_1)}c_2(a_2^{uv})\}\Delta F^{uv(\theta_1)}.$$

where

$$\Delta F_2^{uv(\theta_1)} \equiv F_2(\theta_2^{uv}(\theta_1)) - F_2(\theta_2^{u,v-1}(\theta_1)),$$

and

$$h_2^{uv(\theta_1)} \equiv \int_{\theta_2^{uv-1(\theta_1)}}^{\theta_2^{uv}(\theta_1)} h_2(\theta_2)\, dF_2(\theta_2)/\Delta F_2^{uv(\theta_1)}.$$

This generates the following expression for the principal's expected profit:

$$\pi^d(k_1k_2, k_2 \mid \theta_1) = \sum_{v=1}^{k_2} [B(a^{uv}) - h_1(\theta_1)c_1(a_1^{uv}) - h_2^{uv(\theta_1)}c_2(a_2^{uv})]\Delta F_2^{uv(\theta_1)}.$$

To prove the theorem we establish that for every $\theta_1 \in \Theta_1$,

$$\pi^d(k_1k_2, k_2 \mid \theta_1) \ge \pi^c(k_1k_2, k_1k_2 \mid \theta_1).$$

We shall proceed through a sequence of steps. In what follows, fix $1 \le u \le k_1$ and some type $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$. Given a partition $\{\theta_2^{uv}\}$ and numbers γ and h_1, define

$$W(\{\theta_2^{uv}\}, \gamma \mid h_1) = \sum_{v=1}^{k_2} [B(a^{uv}) - (h_1 + \gamma)c_1(a_1^{uv}) - h_2^{uv}c_2(a_2^{uv})]\,\Delta F_2^{uv}.$$

It follows that given the delegation scheme, type θ_1 of agent 1 selects the partition to maximize $W(\cdot, \gamma(\theta_1) \mid h_1(\theta_1))$, where $\gamma(\theta_1) = \gamma_1^u + b_1(\theta_1) - h_1(\theta_1)$. In contrast, the partition of Θ_2 under centralization will maximize $W(\cdot, \delta(\theta_1) \mid h_1(\theta_1))$, where $\delta(\theta_1) = h_1(\tilde{\theta}_1^u) - h_1(\theta_1)$, and $\tilde{\theta}_1^u$ is defined by

$$h_1(\tilde{\theta}_1^u) = \frac{\int_{\theta_1^{u-1}}^{\theta_1^u} h_1(\theta_1)\, dF_1(\theta_1)}{F_1(\theta_1^u) - F_1(\theta_1^{u-1})}.$$

Let $\{\theta_2^{uv}(h_1, \gamma)\}$ denote the partition that maximizes $W(\cdot, \gamma \mid h_1)$. Defining

$$R(\gamma \mid h_1) = \sum_{v=1}^{k_2} [F_2(\theta_2^{uv}(h_1, \gamma)) - F_2(\theta_2^{u,v-1}(h_1, \gamma))]c_1(a_1^{uv}),$$

it is evident that the resulting profit for the principal is

$$V(\gamma \mid \theta_1) \equiv W(\{\theta_2^{uv}(h_1(\theta_1), \gamma)\}, \gamma \mid h_1(\theta_1)) + \gamma R(\gamma \mid h_1(\theta_1)).$$

In particular, we note that $\pi^c(k_1k_2, k_1k_2 \mid \theta_1) = V(\delta(\theta_1) \mid \theta_1)$, while $\pi^d(k_1k_2, k_2 \mid \theta_1) = V(\gamma(\theta_1) \mid \theta_1)$. To conclude the proof we need the following two technical steps.

STEP 1: *For every h_1, $R(\cdot \mid h_1)$ is nonincreasing.*

STEP 2: *For every $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$, $V(\gamma \mid \theta_1)$ is nondecreasing (nonincreasing) in γ when $\gamma < 0$ ($\gamma > 0$).*

Proof of Step 1: This follows from a standard revealed-preference argument. Take γ and γ' such that $\gamma > \gamma'$. Then

$$W(\{\theta_2^{uv}(h_1, \gamma)\}, \gamma \mid h_1) \geq W(\{\theta_2^{uv}(h_1, \gamma')\}, \gamma \mid h_1),$$

$$W(\{\theta_2^{uv}(h_1, \gamma')\}, \gamma' \mid h_1) \geq W(\{\theta_2^{uv}(h_1, \gamma)\}, \gamma' \mid h_1).$$

Adding these two inequalities, one obtains the claim. □

Proof of Step 2: Define

$$W^*(\gamma \mid h_1) = W(\{\theta_2^{uv}(h_1, \gamma)\}, \gamma \mid h_1)$$

to be the maximized value of $W(\cdot, \gamma \mid h_1)$ over all partitions, so that

$$V(\gamma \mid \theta_1) = W^*(\gamma \mid h_1(\theta_1)) + \gamma R(\gamma \mid h_1(\theta_1)).$$

By a standard envelope argument,

$$W^*(\gamma \mid h_1) = W^*(0 \mid h_1) - \int_0^{\gamma} R(t \mid h_1)\, dt,$$

implying

$$V(\gamma \mid \theta_1) = \int_0^{\gamma} [R(\gamma \mid h_1(\theta_1)) - R(t \mid h_1(\theta_1))]\, dt + V(0 \mid \theta_1).$$

The claim now follows from Step 1. □

Returning to the main proof, suppose the principal sets

$$\gamma_1^u = h_1(\tilde{\theta}_1^u) - b_1(\tilde{\theta}_1^u).$$

Then for $\theta_1 = \tilde{\theta}_1^u$ we have $\delta(\theta_1) = \gamma(\theta_1) = 0$. For $\theta_1 < \tilde{\theta}_1^u$ we have

$$\gamma(\theta_1) = b_1(\theta_1) - h_1(\theta_1) + [h_1(\tilde{\theta}_1^u) - b_1(\tilde{\theta}_1^u)] \geq 0,$$

$$\delta(\theta_1) = h_1(\tilde{\theta}_1^u) - h_1(\theta_1) = \gamma(\theta_1) + [b_1(\tilde{\theta}_1^u) - b_1(\theta_1)] > \gamma(\theta_1).$$

A similar argument shows that when $\theta_1 > \tilde{\theta}_1^u$ then $\delta(\theta_1) < \gamma(\theta_1) < 0$. It follows that $V(\gamma(\theta_1) \mid \theta_1) \geq V(\delta(\theta_1) \mid \theta_1)$, proving that the principal obtains a payoff under delegation which is at least as high as that under centralization.

It remains to show that the delegation mechanism is globally incentive-compatible, i.e., type θ_1 will indeed report m_1^u to the principal. If type θ_1 were to report m_1^w instead of m_1^u, the subsequent subcontracting problem would amount to selecting a partition, which depends on θ_1 and m_1^w, so as to maximize

$$\sum_{v=1}^{k_2} \{B(a^{wv}) - [b_1(\theta_1) + \gamma_1^w] c_1(a_1^{wv}) - h_2^{wv(\theta_1)} c_2(a_2^{wv})\} \Delta F_2^{wv(\theta_1)},$$

with $\{a_1^{wv}, B(a^{wv})\} \in S(m_1^w)$. Let $P(\theta_1, w)$ denote the resulting maximized value of this expression, and let $\{\theta_2^{wv(\theta_1)}\}_{v=1}^{k_2}$ denote an optimal solution to this problem.

Using standard arguments [see Mirrlees (1986) and MMR (1995)], it suffices to demonstrate that the function $\partial P(\theta_1, w)/\partial \theta_1$ is nondecreasing in w for all θ_1. Since

$$\frac{\partial P(\theta_1, w)}{\partial \theta_1} = \sum_{v=1}^{k_2} - b_1'(\theta_1) c_1(a_1^{wv}) \Delta F_2^{wv(\theta_1)}$$

and $\gamma_1^w = h_1(\tilde{\theta}_1^w) - b_1(\tilde{\theta}_1^w)$ is nondecreasing in w, the monotonicity follows from the same argument used in Step 1. □

Proof of Proposition 3: It suffices to show that there exists $u \in \{1, \ldots, k_1\}$ such that $\theta_1^{u-1} < \theta_1^u$, and a set of types $\theta_1 \in (\theta_1^{u-1}, \theta_1^u]$ of positive measure, for which the strict versions of Steps 1 and 2 in the proof of Proposition 2 hold. Specifically, we need to show that for all such θ_1, $V(\gamma \mid \theta_1)$ is *strictly* increasing in γ in a neighborhood of $\gamma(\theta_1)$. In turn this requires that $R(\gamma \mid h_1(\theta_1))$ be strictly decreasing in γ in a neighborhood of $\gamma(\theta_1)$. Let $\tilde{\theta}_1^u$ be defined as in the proof of Proposition 2.

Given arbitrary u, h_1, γ, it follows from straightforward differentiation of $W(\cdot, \gamma \mid h_1)$ that for any $w \in \{1, \ldots, k_2\}$,

$$\frac{\partial}{\partial \theta_2^{uw}} W(\{\theta_2^{uv}\}, \gamma \mid h_1) = f_2(\theta_2^{uw}) n(\theta_2^{uw}; a^{uw}, a^{u,w+1}, h_1, \gamma), \tag{1}$$

where

$$n(\theta_2; a, a', h_1, \gamma) \equiv B(a) - B(a') - (h_1 + \gamma)[c_1(a_1) - c_1(a_1')] - h_2(\theta_2)[c_2(a_2) - c_2(a_2')].$$

Hence we obtain the following properties:

(i) $\partial W / \partial \theta_2^{uw}$ is independent of θ_2^{uv} for any $v \neq w$. Thus, given any pair of partitions over agent 2's type space $\{\theta_2^{uv}\}$ and $\{\tilde{\theta}_2^{uv}\}$, we have

$$W(\{\theta_2^{uv}\}, \gamma \mid h_1) = W(\{\tilde{\theta}_2^{uv}\}, \gamma \mid h_1) + \sum_{v=1}^{k_2} \int_{\tilde{\theta}_2^{uv}}^{\theta_2^{uv}} f_2(x) n(x; a^{uv}, a^{u,v+1}, h_1, \gamma)\, dx. \tag{2}$$

(ii) W is unimodal in θ_2^{uw}, in the sense that $W(\cdot, \gamma \mid h_1)$ is increasing in θ_2^{uw} at any $\theta_2^{uw} < \theta_2^{uw}(h_1, \gamma)$, and decreasing in θ_2^{uw} at any $\theta_2^{uw} > \theta_2^{uw}$ (h_1, γ), strictly so if $a_2^{uw} > a_2^{u,w+1}$. This follows from $a_2^{uw} \geq a_2^{u,w+1}$ by virtue of the incentive compatibility condition for agent 2.

Properties (i) and (ii) together imply:

Step 3: *For arbitrary h_1, γ, and any $\{\tilde{\theta}_2^{uv}\} \neq \{\theta_2^{uv}(h_1, \gamma)\}$,*

$W(\{\theta_2^{uv}(h_1, \gamma)\}, \gamma \mid h_1) > W(\{\tilde{\theta}_2^{uv}\}, \gamma \mid h_1)$

if it is the case that there exists some $v \in \{1, \ldots, k_2\}$ for which

$\theta_2^{uv}(h_1, \gamma) \in \operatorname{int} \Theta_2$ and $a_2^{uv} > a_2^{u,v+1}$.

Step 4: *For some u, $1 \leq u \leq k_1$, such that $\theta_1^{u-1} < \theta_1^u$, there exists a neighborhood N of $\tilde{\theta}_1^u$ such that for any $\theta_1 \in N$, in some neighborhood L of $\gamma(\theta_1)$ it is true that for any $\gamma \in L$, $\theta_2^{uv}(h_1(\theta_1), \gamma)$ lies in the interior of Θ_2, and $a^{uv} \neq a^{u,v+1}$ for some $1 \leq v \leq k_2$.*

Proof: Suppose this is false. Then for every u and every integer n we can select $\theta_1^n \in \Theta_1$ satisfying $|\theta_1^n - \tilde{\theta}_1^u| < 1/n$, and a corresponding sequence $\gamma^{mn} \to \gamma(\theta_1^n)$ as $m \to \infty$, such that for every v with $a^{uv} \neq a^{u,v+1}$, $\theta_2^{uv}(h_1(\theta_1^n), \gamma^{mn})$ is noninterior, i.e., either $\underline{\theta}_2$ or $\overline{\theta}_2$. This implies that for every n, communication with agent 2 is of no value in the problem of maximizing $W(\cdot, \gamma^{nn} \mid h_1(\theta_1^n))$ over all k_2-element partitions over agent 2's type space Θ_2. Since this problem satisfies all the conditions for the theorem of the maximum, taking $n \to \infty$ we obtain a solution to the problem of maximizing $W(\cdot, 0 \mid h_1(\tilde{\theta}_1^u))$ over all partitions. However, this problem reduces exactly to the problem of selecting an optimal centralized contract, contradicting the assumption that communication with agent 2 is valuable under centralization. □

It follows from Step 4 that for every $\theta_1 \in N$ and every $\gamma \in L$ there exists a value of $v \in \{1, \ldots, k_2\}$ such that $\theta_2^{uv}(h_1(\theta_1), \gamma) \in \text{int } \Theta_2$ and $a^{uv} \neq a^{u,v+1}$.

Step 5: *$\theta_2^{uv}(h_1(\theta_1), \gamma) \in \text{int } \Theta_2$ and $a^{uv} \neq a^{u,v+1}$ implies $a_2^{uv} > a_2^{u,v+1}$, $a_1^{uv} \neq a_1^{u,v+1}$, and*

$$\theta_2^{uv}(h_1, \gamma) \neq \theta_2^{uv}(h_1, \gamma') \qquad \text{if} \quad \gamma \neq \gamma'.$$

To prove this, note that by incentive compatibility for agent 2 we have $a_2^{uv} \geq a_2^{u,v+1}$. Now if $a_2^{uv} = a_2^{u,v+1}$, then the production nonseparability condition implies that $a_1^{uv} = a_1^{u,v+1}$, since a^{uv} by definition maximizes $B(a) - h_1^u c_1(a_1) - h_2^v c_2(a_2)$. Hence we contradict the premise that $a^{uv} \neq a^{u,v+1}$, and it must be the case that $a_2^{uv} > a_2^{u,v+1}$.

A similar argument establishes that $a_1^{uv} \neq a_1^{u,v+1}$. Finally, note that $\theta_2^{uv}(h_1(\theta_1), \gamma) \in \text{int } \Theta_2$ implies that $n(\theta_2^{uv}(h_1(\theta_1), \gamma); a^{uv}, a^{u,v+1}, h_1(\theta_1), \gamma) = 0$. For any such v, if it is also true that $a_1^{uv} \neq a_1^{u,v+1}$, then $\gamma \neq \gamma'$ implies

$$\theta_2^{uv}(h_1(\theta_1), \gamma) \neq \theta_2^{uv}(h_1(\theta_1), \gamma'),$$

which completes the proof of Step 5.

Steps 3, 4, and 5 taken together then establish that there exists u with $\theta_1^{u-1} < \theta_1^u$, and a neighborhood N of $\tilde{\theta}_1^u$, such that for any $\theta_1 \in N$ there is a neighborhood L of $\gamma(\theta_1)$ such that for any $\gamma \in L$ and $\gamma' \neq \gamma$ we have

$$W(\{\theta_2^{uv}(h_1(\theta_1), \gamma)\}, \gamma \mid h_1(\theta_1)) > W(\{\theta_2^{uv}(h_1(\theta_1), \gamma')\}, \gamma \mid h_1(\theta_1)).$$

Thus, the first inequality in the proof of Step 1 is strict, and $R(\gamma \mid h_1(\theta_1))$ is strictly decreasing in γ, in a neighborhood of $\gamma(\theta_1)$, for a set of θ_1's of positive measure. This concludes the proof of Proposition 3.

References

Baiman, S., 1991, ''Agency Research in Managerial Accounting: A Second Look,'' *Accounting, Organizations and Society*, 15:4, 341–371.

Baron, D. and D. Besanko, 1992, ''Information, Control and Organizational Structure,'' *Journal of Economics and Management Strategy*, 1, 367–384.

Bolton, P. and M. Dewatripont, 1994, ''The Firm as a Communication Network,'' *Quarterly Journal of Economics*, 109, 809–839.

Che, Y. and D. Hausch, 1996, ''Cooperative Investments and the Value of Contracting: Coase versus Williamson,'' Mimeo, University of Wisconsin.

Dye, R., 1985, ''Costly Contract Contingencies,'' *International Economic Review*, 26:1, 233–250.

Gilbert, R. and M. Riordan, 1995, ''Regulating Complementary Products: A Problem of Institutional Choice,'' *RAND Journal of Economics*, 26:2, 257–276.

Green, J. and J.J. Laffont, 1986, ''Incentive Theory with Data Compression,'' in W. Heller, R. Starr and D. Starett, eds., *Essays in Honor of K.J. Arrow*, vol. 3, Cambridge University Press.

—— and ——, 1987, ''Limited Communication and Incentive Compatibility,'' in T. Groves, R. Radner, and S. Reiter, eds., *Information, Incentives and Economic Mechanisms*, University of Minnesota Press.

—— and ——, 1988, ''Renegotiation and the Form of Efficient Contracts,'' Mimeo, Department of Economics, Harvard University.

Hart, O. and B. Holmstrom, 1987, ''The Theory of Contracts,'' in T.F. Bewley, ed., *Advances in Economic Theory, 5th World Congress of the Econometric Society*, Cambridge University Press, 71–155.

Harris, M., C. Kriebel, and A. Raviv, 1982, ''Asymmetric Information, Incentives and Intrafirm Resource Allocation,'' *Management Science*, 29, 604–620.

Hurwicz, L., 1977, ''On the Dimensional Requirements of Informationally Decentralized Pareto Satisfactory Adjustment Processes,'' in K.J. Arrow and L. Hurwicz, eds., *Studies in Resource Allocation Processes*, Cambridge University Press.

——, 1987, ''On Informational Decentralization and Efficiency in Resource Allocation Mechanisms,'' in S. Reiter, ed., *Studies in Mathematical Economics*, The Mathematical Association of America.

Jordan, J., 1989, ''Accounting Based Divisional Performance Measurement I: Incentives for Profit Maximization,'' *Contemporary Accounting Research*, 6, 903–921.

Kanodia, C., 1993, ''Participative budgets as Coordination and Motivational Devices,'' *Journal of Accounting Research*, 31:2, 172–189.

Laffont, J.J. and J. Tirole, 1993, *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: The MIT Press.

Laffont, J.J. and D. Martimort, 1996, ''Collusion and Delegation,'' Mimeo, Université de Toulouse.

Maskin, E. and J. Tirole, 1990, ''The Principal-Agent Relationship with an Informed Principal: The Case of Private Values,'' *Econometrica*, 58:2, 379–409.

McAfee, P. and J. McMillan, 1995, ''Organizational Diseconomies of Scale,'' *Journal of Economics and Management Strategy*, 4:3, 399–426.

Melumad, N., D. Mookherjee, and S. Reichelstein, 1992, ''A Theory of Responsibility Centers,'' *Journal of Accounting and Economics*, 15, 445–489.

——, 1995, ''Hierarchical Decentralization of Incentive Contracts,'' *Rand Journal of Economics*, 26:4, 654–672.

Mirrlees, J., 1986, ''Optimal Taxation,'' in K.J. Arrow and M. Intrilligator, eds., *Handbook of Mathematical Economics*, vol. III, North Holland.

Mookherjee, D. and S. Reichelstein, 1992, ''Dominant Strategy Implementation of Bayesian Incentive Compatible Allocation Rules,'' *Journal of Economic Theory*, 56:2, 378–399.

——, 1995, ''Incentives and Coordination in Hierarchies,'' Mimeo, Boston University.

——, 1997, ''Budgeting and Hierarchical Control,'' forth-coming, *Journal of Accounting Research*, Spring.

Mount, K. and S. Reiter, 1974, ''The Informational Size of Message Spaces,'' *Journal of Economic Theory*, 8, 161–192.

——, 1995, ''Modelling Bounded Rationality with Modular Networks,'' Mimeo, Northwestern University.

Myerson, R., 1981, ''Optimal Auction Design,'' *Mathematics of Operations Research*, 6, 58–73.

——, 1982, ''Optimal Coordination Mechanisms in Generalized Principal-Agent Problems,'' *Journal of Mathematical Economics*, 10:1, 67–81.

Myerson, R. and M. Satterthwaite, 1983, ''Efficient Bilateral Trading Mechanisms,'' *Journal of Economic Theory*, 29, 265–281.

Qian, Y., 1994, ''Incentives and Loss of Control in Optimal Hierarchy,'' *Review of Economic Studies*, 61:3, 527–544.

Radner, R., 1993, ''The Organization of Decentralized Information Processing,'' *Econometrica*, 60, 1109–1146.

Radner, R. and T. Van Zandt, 1992, ''Information Processing and Returns to Scale,'' *Annales d'Economie et de Statistique*, 25/26, 265–298.

—— and ——, 1995, ''Information Processing and Returns to Scale of a Statistical Decision Problem,'' Mimeo, Princeton University.

Reiter, S., 1996, ''Coordination and the Structure of Firms,'' Mimeo, Northwestern University.

Rogerson, W., 1992, ''Contractual Solutions to the Hold-up Problem,'' *Review of Economic Studies*, 59:4, 777–793.

Williamson, O., 1975, *Markets and Hierarchies*, New York: The Free Press.

——, 1985, *The Economic Institutions of Capitalism*, New York: The Free Press.