Monitoring Teams*

Marina Halac† Ilan Kremer‡ Eyal Winter§

May 11, 2021

Abstract

A principal incentivizes a group of agents to work by choosing a monitoring structure and a scheme of performance-contingent rewards. The monitoring structure partitions the set of agents into monitoring teams, each delivering a signal of joint performance. We show that unlike under partial implementation, the principal always exhausts her monitoring capacity to optimally implement work as a unique outcome. Optimal monitoring teams are homogeneous between them: equally sized and with agents allocated in an anti-assortative fashion. Higher-effort-cost agents receive lower rents, and they tend to be monitored more closely than lower-effort-cost agents when the principal's allocation is constrained.

JEL Codes: D82, D86, L22

Keywords: team incentives, monitoring, contracting with externalities, unique implementation.

---

*We thank Gregorio Curello, Ernesto Dal Bó, Francesc Dilme, Duarte Gonçalves, Sergiu Hart, Mathijs Janssen, Navin Kartik, Elliot Lipnowski, Benny Moldovanu, Daniel Rappoport, Roland Strausz, and various seminar and conference audiences for helpful comments.

†Yale University and CEPR. Email: marina.halac@yale.edu.

‡University of Warwick and The Hebrew University. Email: ikremer@huji.ac.il

§Lancaster University and The Hebrew University. Email: mseyal@mscc.huji.ac.il

1 Introduction

When agents work in a team, incentivizing them can be challenging. Alchian and Demsetz (1972) explain in their famous article: "*With team production it is difficult, solely by observing total output, to either define or determine each individual's contribution [...] [E]ach input owner will have more incentive to shirk when he works as part of a team, than if his performance could be monitored easily or if he did not work as a team*" (Alchian and Demsetz, 1972, pp.779-80). Because monitoring agents' individual performance is costly, the question arises of how monitoring should be optimally structured: "*The costs of metering or ascertaining the marginal products of the team's members is what calls forth new organizations and procedures*" (Alchian and Demsetz, 1972, p.780).

We study a principal who incentivizes a team of agents to exert effort. The principal chooses a monitoring structure that generates information about the agents' performance, together with a scheme of performance-contingent rewards. At one extreme, the principal may only be able to observe a signal of the overall performance of the team; at the other extreme, she may be able to observe a performance signal for each agent. A finer monitoring structure allows the principal to better tie observed performance, and thus each agent's reward, to each agent's effort. When the principal is constrained in her monitoring capacity, how should she allocate monitoring in order to minimize her cost of incentivizing the agents?

The work that followed Alchian and Demsetz (1972) highlighted that, perhaps surprisingly, the monitoring structure does not matter under certain conditions. This is the case in the seminal paper of Holmström (1982) under risk neutrality, and in Picard and Rey (1987) and McAfee and McMillan (1991) featuring adverse selection in addition to moral hazard; they find that a principal can do as well when she observes only the overall team's performance as when she observes each agent's contribution. However, their analyses only deal with how to induce effort as some equilibrium outcome, and they ignore that other equilibria may arise in which some or all agents choose to shirk fearing that other agents will do so too.[1] Based on this observation, Mookherjee and Reichelstein (1992, p.391) suggest that the value of monitoring may be viewed as avoiding such bad equilibrium outcomes.

[1]Such a concern is the focus of a growing literature on unique-implementation schemes that we describe below.

That is the view that we take in this paper. In a setting in which monitoring has no value to the principal under partial implementation, we study the problem of how to optimally ensure effort, namely how to optimally induce the agents to work as a unique equilibrium outcome. We show that, as posited by Alchian and Demsetz (1972), monitoring plays a central role. In fact, not only does greater monitoring allow the principal to uniquely implement work at a lower cost, but the design of monitoring then becomes a key tool for incentive provision. The need to address agents' strategic uncertainty about other agents' effort decisions yields a theory of monitoring with clear implications for organizations.

To focus on monitoring, we consider a stylized setup. Each agent privately chooses whether to work or shirk on an individual task, where working is costly but increases the probability of task completion. The principal monitors the agents' performance by assessing the completion of subsets of tasks. Specifically, we define a monitoring structure as a partition of the set of agents into *monitoring teams*; for each monitoring team, the principal is (only) able to verify whether or not all of its agents have completed their tasks. The principal's cost of monitoring takes the form of a capacity constraint, namely a bound on the number of monitoring teams she can specify.

Our concept of monitoring teams has no a priori revenue implications. The only defining feature of a monitoring team is that the principal can verifiably identify the joint success of its agents, separately from others. In our baseline model, the principal can choose any monitoring partition subject to capacity. In applications, however, there may be production and organizational constraints that limit how tasks can be monitored jointly. For example, take a set of agents tasked with hiring new employees. Their tasks include advertising the job position, selecting applicants for interview, conducting interviews and making offers, and persuading those with offers to accept. The principal can monitor subsets of tasks jointly, but not arbitrarily so; she can separately assess the quality of the interview pool (first two tasks) and the quality of hires given the interview pool (last two tasks), but she cannot jointly evaluate job advertising and offer selection separately from interview selection. Using our framework, we can model these restrictions as constraints on the allocation of agents to monitoring teams and study their implications.

The principal's problem consists of choosing a monitoring structure and an

incentive scheme in order to uniquely implement work at the least possible cost. We solve this problem in two steps. In the first step, we provide a characterization of optimal incentives for any given fixed monitoring structure. The principal solves for an optimal scheme for each monitoring team separately, specifying a bonus payment for each agent conditional on the monitoring team producing good performance. Because such performance depends on the effort choices of all agents in the monitoring team, agents face strategic risk; the principal's bonus offers compensate the agents for this risk in order to guarantee their efforts.

The second step of our solution uses the characterization of optimal incentives to solve for an optimal monitoring structure. We begin by showing that, unlike under partial implementation, the principal's cost of uniquely implementing work does depend on her monitoring capacity. Intuitively, the larger is an agent's monitoring team, the higher is the strategic risk he faces about other agents' effort choices, and thus the higher is the compensation he demands from the principal. Hence, whenever possible, the principal benefits from "splitting" a monitoring team into two. The implication is that every optimal monitoring structure must exhaust the principal's monitoring capacity.

With the number of monitoring teams pinned down by capacity, the question then is how agents should be allocated across them. Our main broad result is that the principal benefits from specifying monitoring teams that are homogeneous with respect to each other. We first show this with regards to size: every optimal monitoring structure consists of monitoring teams of equal size, subject to integer constraints. The reason is that agents' required compensation for strategic risk is convex in the number of other agents with whom their performance is jointly monitored. Put differently, if the principal adds an agent to a monitoring team, she can in principle insure other agents from any additional strategic risk by making it dominant for the new agent to work, but the cost of doing so is exponential in the size of the monitoring team.

If all agents are identical, then equally-sized monitoring teams implies identical monitoring teams. But what if agents differ in their costs of effort? We show that in this case too, it is optimal to make the monitoring teams homogenous with respect to each other, which now requires making them heterogeneous within. Formally, given a monitoring capacity of $\overline{n}$, call the $\overline{n}$ lowest-effort-cost agents rank 1, the next $\overline{n}$ lowest-effort-cost agents rank 2, and so on. Say that a monitoring

structure is *anti-assortative* if no monitoring team contains two agents of the same rank. We show that an anti-assortative monitoring structure is optimal.[2]

The logic behind anti-assortativeness is intuitive. When agents are asymmetric in their effort costs, facing strategic uncertainty about other agents' effort choices is relatively more costly to agents whose effort cost is relatively higher. Hence, the principal benefits from providing higher-cost agents with greater assurance about other agents' effort compared to lower-cost agents. Within each monitoring team, the principal achieves this via the incentive scheme, by offering high enough bonuses to low-cost agents so that their effort is pinned down no matter what higher-cost agents do. The reason an anti-assortative monitoring structure is optimal is that it allows the principal to allocate assurance efficiently not only within but also across monitoring teams. Indeed, given optimal incentives, the key feature of an anti-assortative monitoring structure is that no agent faces higher strategic uncertainty than another agent whose effort cost is comparatively lower, no matter to which monitoring teams the agents belong.

Our results have implications for the design of monitoring and agents' pay in organizations. We find that monitoring is optimally spread evenly across an organization: the principal specifies equally-sized sections whose performance is evaluated separately, with each section containing agents from every rank and thus being diverse within but similar to other sections. This anti-assortativeness prediction is consistent for example with the findings of Adhvaryu et al. (2020), who document negative assortative matching of managers and workers in a large manufacturing firm with production complementarities.[3] Moreover, as a consequence of this allocation, we show that the compensation for strategic risk that agents receive is determined by their ranks. Taking effort costs to be inversely related to skill, this means that the principal offers higher rents to higher-skilled agents compared to lower-skilled agents, not just within each section but across the organization.

While our baseline model of monitoring permits a general characterization, a shortcoming is that it allows for greater flexibility than what may be feasible

[2]Moreover, *every* optimal monitoring structure is anti-assortative absent integer constraints, i.e., if the number of agents divided by $\overline{n}$ is an integer.

[3]Like the agents in our model, these managers and workers are incentivized with bonuses for high team performance. A survey conducted by Adhvaryu et al. (2020) suggests that a driver of the observed allocation is the need to ensure a minimum performance standard in all teams.

in applications. Examples like the hiring one described above suggest that the principal's monitoring structure may be constrained by the production technology and other organizational arrangements. To capture these restrictions, we study a constrained version of our principal's problem. We take the agents to be ordered along a line and we restrict attention to monitoring structures that divide this line into consecutive segments.[4] The principal's unconstrained problem can be viewed as one in which she can choose the order of agents along the line, whereas this order is exogenous in the constrained setting.

The principal's constrained solution is essentially the same as the unconstrained one if all agents are identical. However, if agents differ in their effort costs, then their fixed location along the line limits the principal's ability to specify monitoring teams that are homogenous between them. We show that the main consequence of this constraint is a differential use of monitoring across agents. In particular, the principal now tends to place higher-cost agents in smaller monitoring teams compared to lower-cost agents. As above, the principal benefits from providing greater assurance to agents whose effort cost is relatively higher; when the composition of the monitoring teams is constrained, she achieves this by adjusting their size.

By studying the joint design of monitoring and pay, our analysis sheds light on how these tools are optimally combined to incentivize agents' efforts. We find that when the structure of monitoring is constrained, the principal seeks to tailor monitoring toward higher-cost agents, thus lowering their required compensation for strategic risk and increasing that of lower-cost agents. As a result, the principal tends to incentivize low-skilled agents with close monitoring and low rents, whereas high-skilled agents then enjoy little monitoring and high rents.

Related literature. Our paper relates to three strands of literature. First, we relate to the literature on monitoring in teams. As described above, Alchian and Demsetz (1972) proposed that monitoring is the key reason why providing effort incentives in teams is difficult, but subsequent work challenged their view.[5] We

[4] That is, if a monitoring team contains agents i and $j > i$, then it must also contain all agents whose index is in between.

[5] While Alchian and Demsetz (1972) were particularly interested in the boundaries of the firm, we take their insights as broad motivation to study how monitoring should be optimally organized, and how this organization affects the rewards that agents are offered.

contribute to this literature by highlighting the issue of implementation, clarifying the role of monitoring, and studying its optimal structure for guaranteeing agents' efforts.[6,7]

Second, we relate to the literature on contracting with externalities in multi-agent settings, pioneered by Segal (1999, 2003). This literature focuses on optimal unique-implementation mechanisms as we do. Most closely related to our paper are Winter (2004) and Halac, Lipnowski and Rappoport (2021), both of which examine how to uniquely induce a team of agents to work when only the overall team's success is verifiable. We depart by allowing the principal to obtain finer information about agents' performance via a monitoring structure, and by examining the optimal design of monitoring jointly with that of transfers. As in Winter (2004), we consider public contracts, and as in Halac, Lipnowski and Rappoport (2021), we study agents who may be heterogeneous.[8]

Finally, by examining the allocation of agents to monitoring teams, we relate more broadly to the literature on matching (dating back to Becker, 1973) and on the allocation of agents to productive teams. Most of these papers, however, abstract from incentive provision.[9] Among those that do not, we relate most closely to Kaya and Vereshchagina (2014), Franco, Mitchell and Vereshchagina (2011), and Kambhampati and Segura-Rodriguez (2020).[10] These articles analyze the matching of low- and high-type workers into two-worker teams, taking both revenue and incentive considerations into account. They show that even when revenue maximization calls for positive assortative matching, minimizing the cost of incentives can lead to negative assortative matching depending on functional and parametric assumptions.

[6]Problems of implementation in the provision of team incentives are also discussed in Mookherjee (1984) and, as noted above, Mookherjee and Reichelstein (1992). Baliga (2002) shows that a monitor can help eliminate bad equilibria in an adverse-selection setting.

[7]More tangentially related, there is a literature that studies peer monitoring in teams; see, e.g., Miller (1997), Strausz (1999), Winter (2010), Miller and Rozen (2014), and Gershkov and Winter (2015). Rahman (2012) is concerned with the incentives of the monitor herself.

[8]Heterogeneity is also studied in Bernstein and Winter (2012), Sákovics and Steiner (2012), and Halac, Kremer and Winter (2020). Moriya and Yamashita (2020) extend Winter's (2004) setting by letting the probability of team success depend on an uncertain state.

[9]Meyer (1994) studies task assignments in a dynamic setting with learning about agents' abilities. Prat (2002) analyzes whether a team should be homogenous using team theory. Chade and Eeckhout (2018) examine the matching of agents who differ in their signal informativeness.

[10]More tangentially related, Moldovanu, Sela and Shi (2007) study how to partition heterogeneous agents into status classes according to their efforts, in order to maximize total effort.

2 Model

Setup. Consider a set of tasks, $M = \{1, \ldots, m\}$, each performed by an agent that we index by his task. The tasks contribute to generating an output via a production function that we do not model. Each agent $i \in M$ privately chooses effort $e_i \in \{0, 1\}$, where $e_i = 1$ means "work" and $e_i = 0$ means "shirk." Given e_i, agent i successfully completes his task with probability $p_{e_i} \in (0, 1]$, where $p_0 < p_1$, independent across the agents. Shirking is costless while working entails a cost $c_i > 0$ to agent i. The heterogeneity in effort costs may arise from differences in agents' skills or from differences in the difficulty of their tasks.

A principal monitors the agents' performance by assessing the completion of subsets of tasks. Specifically, we define a *monitoring structure* as a partition π of the set M, where we refer to the parts $S \in \pi$ as *monitoring teams*. A monitoring structure π generates a verifiable signal for each monitoring team $S \in \pi$, which contains all the information then available about the performance of agents $i \in S$. We focus on signals that are binary and supermodular, the latter reflecting that agents' efforts are complementary in yielding a good performance signal. In particular, we take the signal for each monitoring team S to simply indicate whether or not all of its agents have successfully completed their tasks.[11] Thus, given a profile of effort choices $(e_i)_{i \in M}$, S realizes a good signal with probability $\prod_{i \in S} p_{e_i}$ and a bad signal otherwise.

The principal's cost of monitoring takes the form of a capacity constraint: the number n of monitoring teams that she can specify in a monitoring structure $\pi = \{S_1, \ldots, S_n\}$ is capped by a number $\overline{n} \geq 1$. If $\overline{n} = m$, each individual task can be monitored, so this capacity constraint is non-binding. If $\overline{n} = 1$, the only available signal corresponds to the whole set M of tasks, so the design of monitoring is moot. We are interested in studying the principal's optimal monitoring structure when $\overline{n}$ is in between these two extremes.

Principal's problem. Fix an integer $1 \leq \overline{n} \leq m$ and let $\Pi(\overline{n})$ be the set of all partitions π of M such that the number of parts is $n \leq \overline{n}$. The principal chooses a monitoring structure $\pi \in \Pi(\overline{n})$, which generates verifiable signals as

[11] This signal structure corresponds to the production function in Kremer's (1993) O-ring theory and in the benchmark model of Winter (2004). See Section 5.3 for further discussion.

described above. Additionally, for each monitoring team $S \in \pi$ and agent $i \in S$, the principal offers the agent an incentive contract to induce him to work.

We consider public contracts and assume that agents are protected by limited liability, so any payments from the principal must be nonnegative. Note that for each monitoring team $S \in \pi$, the principal can only verify whether or not all agents $i \in S$ have completed their tasks, which depends on these agents' effort choices only and not those of other agents $j \notin S$. Hence, without loss, the contract for agent $i \in S$ simply specifies a bonus b_i for the agent if S produces a good signal and zero payment otherwise.

The bonus offers $b_S := (b_i)_{i \in S}$ define a simultaneous game among the agents in S. In this game, each agent $i \in S$ chooses whether to work or shirk, with his payoff being equal to his expected bonus payment, minus his effort cost c_i if he works. The principal wishes to ensure effort at the least possible cost, that is, to specify a least-cost incentive scheme such that all agents working is the unique Nash equilibrium of the game induced in each $S \in \pi$. A technical issue is that the set of schemes b_S that ensure effort in S is open (since b_i takes continuous values). We resolve this by requiring the principal's scheme to induce a unique equilibrium only once the bonus offers are increased by any positive amount.[12] Formally, say that $(b_S)_{S \in \pi}$ *uniquely implements work (UIW)* if, for each $S \in \pi$ and every $\varepsilon > 0$, all agents $i \in S$ working is the unique Nash equilibrium of the game defined by $b_S + \varepsilon$.

Let $f(S) := p_1^{|S|}$ denote the probability that monitoring team S realizes a good signal conditional on all agents $i \in S$ working. The principal's problem is to choose a monitoring structure and an incentive scheme to minimize her total incentive cost, subject to the number of monitoring teams not exceeding $\overline{n}$ and to uniquely implementing work in each of them:

$$\min_{\pi \in \Pi(\overline{n}), (b_S)_{S \in \pi}} \sum_{S \in \pi, i \in S} f(S) b_i \qquad \text{(P)}$$

$$\text{subject to } (b_S)_{S \in \pi} \text{ UIW.}$$

As a remark, we observe that in our problem the requirement of unique im-

[12] This is equivalent to assuming that agents work when indifferent between working and shirking given their conjectures of others agents' behavior.

plementation in Nash equilibria is equivalent to requiring a unique rationalizable outcome. This follows from the performance signals being supermodular and the agents being protected by limited liability. Given these, for any monitoring structure $\pi \in \Pi(\overline{n})$ and incentive scheme $(b_S)_{S\in\pi}$ that the principal chooses, the game induced in each $S \in \pi$ is a supermodular game, and thus the results of Milgrom and Roberts (1990) apply. An implication is that our analysis does not rely on strong assumptions about agents' ability to predict others' behavior; it only relies on agents being rational and this rationality being common knowledge.

Monitoring constraints. Our baseline model places no constraints on the principal's monitoring partition other than her monitoring capacity. In applications, however, there may be additional restrictions arising from the production technology and organizational arrangements which limit how tasks can be grouped to be monitored jointly. Using our framework, we can capture these as constraints on the principal's allocation of agents to monitoring teams. We introduce such constraints and study their implications in Section 4.3.

3 Partial implementation benchmark

Before we solve the principal's problem in (P), we consider a relaxed version of this problem which ignores the unique implementation constraint. We show that if the principal only seeks to implement work as a Nash equilibrium outcome, not necessarily the unique one, then the design of monitoring becomes trivial: the principal's value is independent of the monitoring capacity $\overline{n}$ and the monitoring structure $\pi \in \Pi(\overline{n})$.

Suppose that the principal specifies a monitoring structure $\pi \in \Pi(\overline{n})$ and, for each monitoring team $S \in \pi$, she only wishes to incentivize all agents $i \in S$ working as some Nash equilibrium of the induced game. In this case, for each $S \in \pi$, the principal's bonus offers need only ensure that each agent $i \in S$ prefers to work rather than shirk when he conjectures that all other agents in S will work. Formally, for each $S \in \pi$ and each $i \in S$, the bonus b_i must satisfy

$$f(S)b_i - c_i \geq f(S)\frac{p_0}{p_1}b_i.$$

The principal's optimal bonus offers make these incentive constraints hold with equality:

$$b_i^{NE} = \frac{p_1}{p_1 - p_0} \frac{c_i}{f(S)}. \tag{1}$$

Substituting the bonuses in the principal's objective in (P), we obtain that the principal's incentive cost is equal to

$$\sum_{S \in \pi, i \in S} \frac{p_1}{p_1 - p_0} c_i. \tag{2}$$

Since this expression is independent of π, the next result follows immediately.

Proposition 1. *Under partial implementation, the principal's cost of incentives is independent of the monitoring capacity and the monitoring structure.*

As discussed in the Introduction, this result is consistent with a previous literature, most notably the seminal paper of Holmström (1982) in the case that agents are risk neutral as in our model.[13] Proposition 1 says that monitoring agents' individual performance has no value to the principal. Even if the principal could costlessly monitor the completion of each individual task, her cost of incentivizing the agents would be the same as when she is only able to verify whether or not the whole set of tasks have been successfully completed.

The takeaway from this section is that partial implementation is not well-suited to study the principal's monitoring problem. Focusing on partial implementation not only suffers from the standard selection criticism—the fact that the principal may be unable to coordinate agents to play her preferred equilibrium when multiple equilibria exist—but it also fails to capture the incentive benefits of using finer monitoring structures. This benchmark in hand, we turn to our analysis of optimal monitoring under unique implementation.

4 Optimal monitoring structure

To solve the principal's problem in (P), we proceed as follows. First, in Section 4.1, we provide a characterization of optimal incentives for any given fixed monitoring

[13] While in our setting this result relies on the signal structure that we have assumed, we view it as a meaningful benchmark precisely because the literature has shown it to hold under more general conditions.

structure. Next, in Section 4.2, we use this characterization to solve for an optimal monitoring structure. Finally, in Section 4.3, we introduce additional constraints on the monitoring allocation and study their implications.

4.1 Incentives

Fix a monitoring structure $\pi \in \Pi(\overline{n})$ and consider an optimal incentive scheme $\{b_S\}_{S\in\pi}$. Note that given π fixed, the principal solves for an optimal scheme b_S for each monitoring team $S \in \pi$ separately, in order to minimize the cost of ensuring work in S. This problem is related to the team problems studied in Winter (2004) and Halac, Lipnowski and Rappoport (2021).

We begin by showing that if an incentive scheme uniquely implements work in monitoring team S, then it must make it iteratively dominant for each agent $i \in S$ to work:

Lemma 1. *Fix a monitoring structure $\pi \in \Pi(\overline{n})$ and suppose b_S uniquely implements work in $S \in \pi$. Then there exists a permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S such that, for each $j \in \{1, \ldots, |S|\}$, agent i_{Sj} is willing to work if agents $(i_{S1}, \ldots, i_{Sj-1})$ work, no matter what agents $(i_{Sj+1}, \ldots, i_{S|S|})$ do.*

The logic is simple. If a scheme b_S uniquely implements work in S, then there must be an agent $i_{S1} \in S$ who is willing to work under b_S when the other agents in S shirk. Moreover, by supermodularity of the signal structure, this agent is thus willing to work no matter what the other agents in S do. Proceeding by induction delivers the result in Lemma 1.

We next use this result to derive a characterization of optimal incentives for any given monitoring team S. An optimal incentive scheme for S specifies some permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S and a bonus $b_{i_{Sj}}$ for each agent $i_{Sj} \in S$ satisfying the criterion in Lemma 1. Let us first fix a permutation $(i_{S1}, \ldots, i_{S|S|})$ and solve for optimal bonuses given this permutation. By the lemma, agent i_{Sj} must be willing to work if agents $(i_{S1}, \ldots, i_{Sj-1})$ work, no matter the rest. By supermodularity, this is true if and only if agent i_{Sj} is willing to work when agents $(i_{S1}, \ldots, i_{Sj-1})$ work and agents $(i_{Sj+1}, \ldots, i_{S|S|})$ shirk. Hence, for each $j \in \{1, \ldots, |S|\}$, the bonus $b_{i_{Sj}}$ must satisfy

$$p_1^j p_0^{|S|-j} b_{i_{Sj}} - c_{i_{Sj}} \geq p_1^{j-1} p_0^{|S|-j+1} b_{i_{Sj}}.$$

The principal's optimal bonus offers make these incentive constraints hold with equality. Rearranging terms, we thus obtain

$$b^*_{i_{Sj}} = \frac{p_1}{p_1 - p_0} \frac{c_{i_{Sj}}}{f(S)} \left(\frac{p_1}{p_0}\right)^{|S|-j}.$$

It will be useful to define the *compensation factor*

$$r_{Sj} := \left(\frac{p_1}{p_0}\right)^{|S|-j}, \tag{3}$$

which allows us to rewrite the optimal bonus for each $j \in \{1, \ldots, |S|\}$ as

$$b^*_{i_{Sj}} = \frac{p_1}{p_1 - p_0} \frac{c_{i_{Sj}}}{f(S)} r_{Sj}.$$

Finally, having characterized the optimal bonus schedule for any given permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S, we now solve for an optimal permutation $(i^*_{S1}, \ldots, i^*_{S|S|})$. Observe that $b^*_{i_{Sj}}$ is supermodular in $c_{i_{Sj}}$ and r_{Sj}, and the compensation factor r_{Sj} is decreasing in j, i.e., it is lower for agents who are placed later in the permutation. Since the principal wishes to minimize the sum of bonus payments, it follows that an optimal permutation of the agents in S orders the agents by increasing cost of effort: $c_{i^*_{Sj}} \leq c_{i^*_{Sj+1}}$ for each $j \in \{1, \ldots, |S| - 1\}$.

Proposition 2 summarizes our findings.

Proposition 2. *Fix a monitoring structure $\pi \in \Pi(\overline{n})$ and suppose $\{b^*_S\}_{S \in \pi}$ is optimal given π. Then for each $S \in \pi$, there exists a permutation $(i^*_{S1}, \ldots, i^*_{S|S|})$ of the agents in S such that $c_{i^*_{S1}} \leq \ldots \leq c_{i^*_{S|S|}}$ and, for each $j \in \{1, \ldots, |S|\}$,*

$$b^*_{i^*_{Sj}} = \frac{p_1}{p_1 - p_0} \frac{c_{i^*_{Sj}}}{f(S)} r_{Sj}. \tag{4}$$

It is instructive to compare the optimal bonuses derived in equation (4) under unique implementation with the bonuses derived in equation (1) under partial implementation. We find that for any given monitoring team S, the principal must offer the agents in S higher bonuses than in the partial implementation benchmark in order to guarantee their efforts. This is reflected in the compensation factors $r_{Sj} \geq 1$ which appear in the bonuses $b^*_{i^*_{Sj}}$ above but not in equation (1). To

induce work as a unique outcome, the principal must compensate the agents for the strategic risk that they face about the effort choices of other agents in S. This risk arises because the principal only observes a signal for the joint performance of all agents in a monitoring team. The possibility that other agents in S may shirk implies a lower marginal effect of each agent's effort on joint performance, and therefore a higher performance bonus required by each agent in S to work.

Proposition 2 shows that the compensations that agents $i \in S$ receive vary according to the permutation of these agents that is specified by the principal's incentive scheme. An agent $i \in S$ faces higher strategic risk and thus demands a higher compensation the earlier he is placed in the permutation. This in turn explains why, for any given monitoring team S, an optimal permutation orders the agents in S by increasing cost of effort. Intuitively, since facing strategic uncertainty about other agents' efforts is more costly to higher-cost agents, the principal benefits from providing higher-cost agents with assurance that lower-cost agents will work. Hence, in any given monitoring team, lower-cost agents are placed earlier in the permutation and receive a higher compensation factor than higher-cost agents. This means that lower-cost agents are offered higher markups than higher-cost agents, both relative to their effort costs, $b^*_{i^*_{Sj}}/c_{i^*_{Sj}}$, as well as relative to the partial-implementation bonuses, $b^*_{i^*_{Sj}}/b^{NE}_{i^*_{Sj}}$.

4.2 Monitoring structure

We now proceed to study the principal's optimal monitoring structure. Using the characterization of optimal incentives in Proposition 2, we can rewrite the principal's incentive cost and thus simplify the principal's problem in (P). Substituting with the bonuses in (4), the principal's problem reduces to choosing a monitoring structure $\pi \in \Pi(\overline{n})$ in order to minimize

$$\sum_{S \in \pi, j \in \{1,\ldots,|S|\}} \frac{p_1}{p_1 - p_0}\, c_{i^*_{Sj}}\, r_{Sj}, \tag{5}$$

where, for each $S \in \pi$ and each $j \in \{1, \ldots, |S|\}$, $i^*_{Sj} \in S$ and $c_{i^*_{S1}} \leq \ldots \leq c_{i^*_{S|S|}}$.

The principal's incentive cost depends on the monitoring partition $\pi = \{S_1, \ldots, S_n\}$ via the compensation factors r_{Sj} which multiply each term in the sum in (5).[14]

[14] In fact, if these factors were equal to 1 for all $S \in \pi$ and $j \in \{1, \ldots, |S|\}$, then the principal's

Recall from the definition in (3) that $r_{Sj} \geq 1$ for all $S \in \pi$ and $j \in \{1, \ldots, |S|\}$, strictly increasing in $|S| - j$. This implies that the principal's incentive cost is strictly increasing in the size of the monitoring teams, yielding the following result:

Proposition 3. *If $\pi^* = \{S_1^*, \ldots, S_{n^*}^*\}$ is an optimal monitoring structure, then $n^* = \overline{n}$ and the principal's cost of incentives is strictly decreasing in the monitoring capacity.*

This result contrasts with that in Proposition 1. Unlike under partial implementation, we find that the principal's cost of uniquely implementing work does depend on her monitoring capacity. As discussed in the Introduction, this result therefore supports Alchian and Demsetz's (1972) view on the value of monitoring, as well as Mookherjee and Reichelstein's (1992, p.391) suggestion that this value may arise from the need to exclude bad equilibrium outcomes.

The intuition for Proposition 3 is related to our discussion in Section 4.1. To ensure an agent's effort, the principal must compensate the agent for the strategic risk that he faces about the effort choices of other agents in his monitoring team. The principal can reduce the agent's strategic risk (and thus his required compensation) by reducing the size of his monitoring team, as that makes the agent's pay dependent on a fewer number of other agents. Naturally, if $n = \overline{n}$, then reducing the size of a monitoring team would imply increasing the size of another monitoring team. But so long as $n < \overline{n}$, the principal can strictly lower her cost of incentives by splitting a monitoring team into two. Therefore, any optimal monitoring structure must exhaust the principal's monitoring capacity by setting $n = \overline{n}$.

The next two results describe how the principal optimally partitions the set of agents into $\overline{n}$ monitoring teams. We first show that the principal benefits from specifying monitoring teams of equal size (subject to integer constraints). The reason is that the compensation factor r_{Sj} that multiplies each term in (5) is convex in $|S| - j$, which means that the principal's cost of uniquely implementing work in a monitoring team S is convex in its size.

Proposition 4. *If π^* is an optimal monitoring structure, then $|S| - |S'| \leq 1$ for every $S, S' \in \pi^*$.*

cost would coincide with her cost under partial implementation given by (2), and would thus be independent of the monitoring structure.

To see the logic, suppose by contradiction that a monitoring structure π is optimal with $S, S' \in \pi$ and $|S| - |S'| > 1$. Let i be a lowest-effort-cost agent in monitoring team S. We show that the principal's incentive cost can be strictly reduced by moving agent i from S to S'. Observe that by Proposition 2, an optimal incentive scheme for S makes it dominant for agent i to work (i.e., it places i first in the corresponding permutation for S). It thus suffices to show that the principal's incentive cost strictly declines when we move i to S' and continue to make it dominant for i to work in S'. In fact, in this case, moving agent i from S to S' does not affect the strategic risk faced by any other agent, so the move benefits the principal provided that it strictly reduces the strategic risk faced by agent i. The latter is true under $|S| - |S'| > 1$.

Proposition 4 gives a full characterization of the principal's optimal monitoring structure if all agents are symmetric, i.e., if $c_i = c$ for all $i \in M$. In this case, the result that all monitoring teams are equally-sized implies that they are all identical (again, subject to integer constraints).

Suppose instead that agents differ in their costs of effort. The following definition will be useful to describe the principal's solution.

Definition 1. *Let Φ be the set of permutations $\phi = (i_1, \ldots, i_m)$ of M with $c_{i_1} \leq \ldots \leq c_{i_m}$. Given $\phi \in \Phi$, say the first $\overline{n}$ agents in ϕ are rank 1, the next $\overline{n}$ agents are rank 2, and so on. A monitoring structure $\pi \in \Pi(\overline{n})$ is* anti-assortative *if there is $\phi \in \Phi$ such that no monitoring team in π contains two agents of the same rank.*

We can construct an anti-assortative monitoring structure using the following simple procedure. We first define ranks as described in Definition 1: we take a permutation $(i_1, \ldots, i_m)$ that orders the agents by increasing cost of effort, and we assign rank 1 to the $\overline{n}$ lowest-cost agents in the permutation, rank 2 to the next $\overline{n}$ lowest-cost agents, and so on.[15] Next, we assign each of the rank-1 agents to a different monitoring team; for example, agent i_1 to monitoring team S_1, agent i_2 to monitoring team S_2, and so on until agent $i_{\overline{n}}$. We then assign each of the rank-2 agents to a different monitoring team; for example, agent $i_{\overline{n}+1}$ to monitoring team S_1, agent $i_{\overline{n}+2}$ to monitoring team S_2, and so on until agent $i_{2\overline{n}}$. Repeating this

[15] Let z be the lowest integer such that $z \geq m/\overline{n}$. Then there are z ranks, each of the first $z-1$ containing $\overline{n}$ agents and rank z containing $m - (z-1)\overline{n}$ agents.

with each subsequent rank until all agents have been assigned to a monitoring team yields an anti-assortative monitoring structure $\pi = \{S_1, \ldots, S_{\overline{n}}\}$.

We obtain:

Proposition 5. *There exists an optimal monitoring structure that is anti-assortative.*

An optimal monitoring structure π^* specifies monitoring teams that are homogeneous between them: every $S \in \pi^*$ has one agent from each of the effort-cost ranks defined above (up to integer constraints). The intuition is that this partition permits an optimal provision of assurance. Because facing strategic uncertainty about other agents' effort choices is more costly to agents with higher effort costs, the principal benefits from providing higher-cost agents with greater assurance relative to lower-cost agents. The characterization of optimal incentives in Proposition 2 applies this logic within each monitoring team, and the characterization of optimal monitoring in Proposition 5 extends this logic across monitoring teams. Indeed, given optimal incentives, the key feature of an anti-assortative structure is that no agent faces higher strategic risk than another agent whose effort cost is comparatively lower, no matter to which monitoring teams the agents belong.

The proof of Proposition 5 is constructive. Suppose that π is an optimal monitoring structure. If it is not anti-assortative, there must be monitoring teams $S, S' \in \pi$ and agents $i \in S$ and $i' \in S'$ such that agent i has a higher cost of effort than agent i' yet he faces higher strategic risk in S than i' does in S'.[16] We show that the principal's incentive cost can then be weakly reduced by swapping agents i and i', and performing this perturbation on every other such two agents delivers a monitoring structure that is anti-assortative. Moreover, we also show that absent integer constraints (i.e., if $m/\overline{n}$ is an integer), the perturbation reduces the principal's incentive cost strictly, implying that every optimal monitoring structure must be anti-assortative in this case.

There are two noteworthy implications that follow from the result in Proposition 5. The first implication concerns the structure of monitoring itself. We find that by specifying monitoring teams that are homogeneous with respect to each other, the principal's solution specifies monitoring teams that are heterogeneous within. Intuitively, the principal divides the organization into equally-sized

[16] That is, agent i is optimally placed earlier in the permutation for S compared to agent i' in the permutation for S'.

sections whose performance is evaluated separately, with each section containing agents from every rank and thus being diverse within but similar to other sections.

Our anti-assortativeness prediction is in line with recent empirical work by Adhvaryu et al. (2020), who document negatively assortative matched teams in a large readymade garment manufacturer in India. Their teams consist of a manager and a set of workers, all incentivized with bonuses for high team performance. The authors find that revenue considerations would call for positive assortative matching of managers to workers. Instead, according to a survey of managers, negative assortative matching is used to best ensure a minimum performance standard in all teams.

The second implication of our results concerns agents' pay. We find that under the principal's optimal monitoring structure, agents' compensation factors are determined by their ranks, with rank-1 agents receiving the highest factor, rank-2 agents the second-highest factor, and so on. Consequently, lower-cost agents are offered higher markups than higher-cost agents, not only within each monitoring team but across all agents $i \in M$.

Example. We close this section with a simple example that illustrates our results. Take a set $M = \{1, 2, 3, 4\}$ and a monitoring capacity of $\overline{n} = 2$. Assume $c_1 = c_2 = c_L$ and $c_3 = c_4 = c_H$ for $c_L < c_H$. By Proposition 2, this implies that optimal incentives for any given monitoring team can be specified under the identity permutation of its agents.

An anti-assortative monitoring structure in this example is $\pi^* = \{\{1, 3\}, \{2, 4\}\}$, as we illustrate in Figure 1. Using (5), the principal's incentive cost under this monitoring structure is equal to

$$\frac{p_1}{p_1 - p_0} 2 \left(c_L \frac{p_1}{p_0} + c_H \right).$$

To illustrate our result in Proposition 3, we can compare this incentive cost under π^* with the one that would result if the principal only monitors the joint performance of the whole set M of agents. The latter is equal to

$$\frac{p_1}{p_1 - p_0} \left[c_L \left(\frac{p_1}{p_0} \right)^3 + c_L \left(\frac{p_1}{p_0} \right)^2 + c_H \left(\frac{p_1}{p_0} + 1 \right) \right],$$

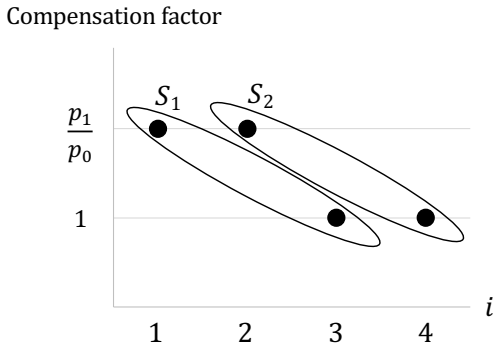

Figure 1: Optimal monitoring structure and compensation factors for $M = \{1, 2, 3, 4\}$ and $\overline{n} = 2$ with $c_1 = c_2 = c_L < c_H = c_3 = c_4$.

and is thus strictly higher than that under π^*.

To illustrate our result in Proposition 4, it suffices to compare the incentive cost under π^* with the one that would result from the monitoring structure $\{\{1, 2, 3\}, \{4\}\}$. The latter is equal to

$$\frac{p_1}{p_1 - p_0}\left[c_L\left(\frac{p_1}{p_0}\right)^2 + c_L\frac{p_1}{p_0} + 2c_H\right], \tag{6}$$

and is thus also strictly higher than the incentive cost under π^*.

Finally, to illustrate our result in Proposition 5, it suffices to compare the incentive cost under π^* with the one that would result from the monitoring structure $\{\{1, 2\}, \{3, 4\}\}$. The latter is equal to

$$\frac{p_1}{p_1 - p_0}(c_L + c_H)\left(\frac{p_1}{p_0} + 1\right), \tag{7}$$

and is thus also strictly higher than the incentive cost under π^*.

4.3 Monitoring constraints

We have so far assumed that the principal's choice of a monitoring partition is only constrained by her monitoring capacity. As illustrated by the hiring example in the Introduction, in reality there may be additional constraints arising from the production technology and other organizational arrangements. These features

determine the nature of agents' tasks and how they are combined to produce output. Thus, while the principal may be able to isolate the performance of certain subsets of tasks, including possibly each individual task, she may not be able to group tasks in an arbitrary manner to monitor their performance jointly. In this section, we introduce these constraints on the principal's monitoring problem.

To represent the monitoring constraints, consider the following simple formulation. We take the agents' tasks to be exogenously ordered along the line $(1, \ldots, m)$, and we restrict attention to monitoring structures that divide this line into consecutive segments. Formally, let $\Pi^C(\overline{n})$ be the set of all partitions π of $M = \{1, \ldots, m\}$ such that (i) the number of parts is $n \leq \overline{n}$, and (ii) for each $S \in \pi$ and $i < j < k$, if $i, k \in S$, then $j \in S$. Note that this set is smaller than the set $\Pi(\overline{n})$ considered in our baseline model; the latter ignores condition (ii) which constrains how agents can be grouped into monitoring teams. Together with condition (ii), the order of agents' tasks $(1, \ldots, m)$ represents the fixed production and organizational arrangements that the principal's monitoring allocation must respect. Our notation for a partition $\pi = \{S_1, \ldots, S_n\} \in \Pi^C(\overline{n})$ will index the parts according to this order; that is, for each $k \in \{1, \ldots, n-1\}$, we take monitoring team S_k to contain lower-indexed agents than monitoring team S_{k+1}.

The principal's constrained problem coincides with her unconstrained problem in (P) except for the fact that she can only choose monitoring structures $\pi \in \Pi^C(\overline{n})$:

$$\min_{\pi \in \Pi^C(\overline{n}), (b_S)_{S \in \pi}} \sum_{S \in \pi, i \in S} f(S) b_i \qquad \text{(P-constrained)}$$
$$\text{subject to } (b_S)_{S \in \pi} \text{ UIW.}$$

To solve this program, we can proceed in an analogous manner as we did to solve the unconstrained problem. Recall that our characterization of optimal incentives in Proposition 2 applies to any given monitoring structure. We can thus use that characterization to substitute for the bonuses in (P-constrained) and simplify the program. The principal's constrained problem reduces to choosing a monitoring structure $\pi \in \Pi^C(\overline{n})$ in order to minimize her cost of incentives given by expression (5).

Observe that Proposition 3 continues to hold in this setting. Plainly, the prin-

cipal can always split a monitoring team into two and reduce the strategic risk agents face, so any optimal monitoring structure $\pi^* = \{S_1^*, \ldots, S_{n^*}^*\}$ must exhaust the principal's monitoring capacity by setting $n^* = \overline{n}$.[17] Moreover, note that if $c_i = c$ for all $i \in M$, then there is an optimal monitoring structure in the unconstrained problem which is feasible in the constrained problem, implying that such a monitoring structure is also optimal in the constrained problem. Hence, the principal's solution is essentially unchanged by the additional monitoring constraints when all agents are symmetric.

Proposition 6. *Suppose $\pi^* = \{S_1^*, \ldots, S_{n^*}^*\}$ is an optimal monitoring structure in the constrained problem. Then $n^* = \overline{n}$. Moreover, if $c_i = c$ for all $i \in M$, then $|S| - |S'| \leq 1$ for every $S, S' \in \pi^*$.*

Things however can be different when agents are asymmetric in their costs of effort. In fact, the restriction in (P-constrained) is that the assignment of agents to monitoring teams is now constrained by the agents' location on the line $(1, \ldots, m)$. Put differently, the principal cannot choose the allocation of effort costs along $(1, \ldots, m)$, and therefore an optimal monitoring structure will depend on how such costs vary along this order. Proposition 7 considers two configurations, one in which effort costs can only be either low or high, and one in which effort costs are monotonically ordered along $(1, \ldots, m)$. The latter may arise in applications if the principal's ability to monitor two agents jointly is greater when these agents bear similar effort costs, for example because they work on similar tasks.

Proposition 7. *Suppose $\pi^* = \{S_1^*, \ldots, S_{\overline{n}}^*\}$ is an optimal monitoring structure in the constrained problem.*

1. *If $c_i \in \{c_L, c_H\}$ for all $i \in M$, then $|\{i \in S : c_i = c_H\}| \leq |\{i \in S' : c_i = c_H\}|$ for every adjacent $S, S' \in \pi^*$ with $|S| - |S'| > 1$.*

2. *If $c_i < c_{i+1}$ for each $i \in M$, then $|S_k^*| \geq |S_{k+1}^*|$ for each $1 \leq k < \overline{n}$.*

Proposition 7 shows that the principal tends to place agents with higher effort costs in smaller monitoring teams compared to agents with lower effort costs. This result contrasts with Proposition 4 in the unconstrained setting, where all

[17] The proof of Proposition 3 in Appendix A applies to the constrained problem by letting $\iota \in S_k$ be such that $\iota + 1 \notin S_k$.

agents are assigned to monitoring teams of equal size. Therefore, we find that the main consequence of the additional monitoring constraints is a differential use of monitoring across agents: the principal now optimally tailors monitoring toward higher-cost agents.

The intuition for this result is familiar by now. Since compensating high-cost agents for strategic risk is more expensive than compensating low-cost agents, the principal benefits from providing high-cost agents with greater assurance. In the unconstrained problem, the principal achieves this by specifying an anti-assortative monitoring structure, matching high-cost agents with low-cost ones whose effort can be more cheaply pinned down, and thus delivering monitoring teams that are homogeneous between them. In the constrained problem, the principal also seeks to form monitoring teams whose composition of effort costs is similar to each other. However, since the composition of the monitoring teams is now constrained, the principal needs to adjust their size. When unable to match agents flexibly, the principal addresses the strategic risk of high-cost agents by making their monitoring teams relatively smaller.

The implications for agents' pay are immediate. Note that the principal's monitoring structure determines the extent to which each agent's effort is incentivized with monitoring versus markup. The results above suggest that high-cost agents in an organization will tend to be incentivized with close monitoring and low rents, whereas low-cost agents will then enjoy little monitoring and high rents.

Example. We return to the example described in Section 4.2 with $M = \{1, 2, 3, 4\}$, $\overline{n} = 2$, $c_1 = c_2 = c_L$, and $c_3 = c_4 = c_H$ for $c_L < c_H$. Recall that in the unconstrained problem, an optimal monitoring structure is given by $\{\{1, 3\}, \{2, 4\}\}$, as illustrated in Figure 1. This monitoring structure is not feasible in the constrained problem.

There are two relevant monitoring structures to consider in the constrained setting, $\{\{1, 2\}, \{3, 4\}\}$ and $\{\{1, 2, 3\}, \{4\}\}$. Figure 2 depicts these monitoring structures and the resulting compensation factors for each of the agents. Observe that compared to $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 2, 3\}, \{4\}\}$ requires the principal to pay higher compensation factors to the low-cost agents 1 and 2, while allowing her to pay lower compensation factors to the high-cost agents 3 and 4.

The principal's incentive cost under $\{\{1, 2\}, \{3, 4\}\}$ is given by (7), and her

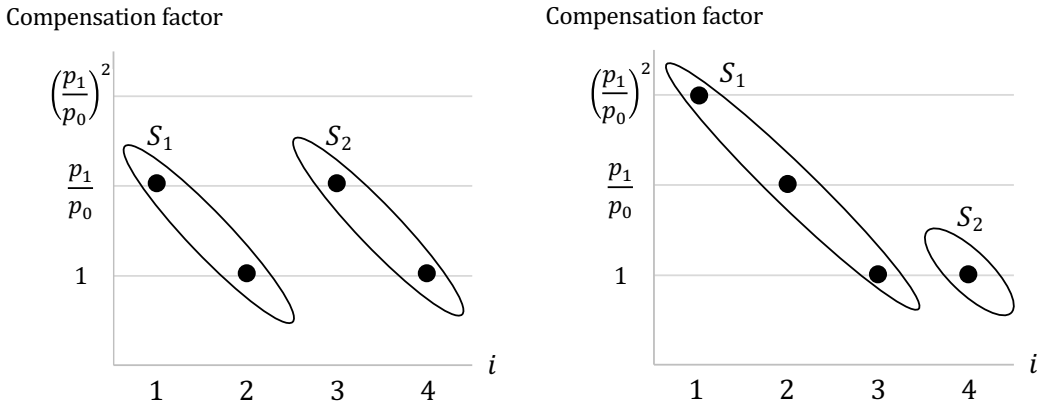

Figure 2: Monitoring structure and compensation factors for $M = \{1, 2, 3, 4\}$ and $\overline{n} = 2$. Taking $c_1 = c_2 = c_L < c_H = c_3 = c_4$, the left panel depicts the principal's constrained solution under $\left(\frac{p_1}{p_0} + 1\right) c_L > c_H$, the right panel under $\left(\frac{p_1}{p_0} + 1\right) c_L < c_H$.

incentive cost under $\{\{1, 2, 3\}, \{4\}\}$ is given by (6). As shown in Section 4.2, both of these monitoring structures perform strictly worse than the principal's unconstrained solution $\{\{1, 3\}, \{2, 4\}\}$. But which one is optimal in the constrained setting? Comparing expressions (6) and (7) yields that $\{\{1, 2, 3\}, \{4\}\}$ is optimal if c_H is sufficiently higher than c_L, namely

$$\left(\frac{p_1}{p_0} + 1\right) c_L < c_H, \tag{8}$$

whereas $\{\{1, 2\}, \{3, 4\}\}$ is optimal when the opposite is true. As depicted in the right panel of Figure 2, when (8) holds, the principal incentivizes the high-cost agents with closer monitoring (on average) and lower markups compared to the low-cost agents.

5 Discussion

In this section, we discuss several possible extensions of the problem that we have studied.

5.1 Task completion

We have studied agents who may differ only in their cost of effort. Suppose now that agents may differ in the probability with which they complete their tasks conditional on effort. For each $i \in M$, let p_i be the probability that agent i completes his task if he works; if the agent shirks, his probability of task completion is p_0, with $0 < p_0 < p_i \leq 1$ for all $i \in M$. For simplicity, assume that agents are symmetric in their effort costs, i.e., $c_i = c$ for all $i \in M$.

The principal's problem is given by (P) once we redefine $f(S) := \prod_{i \in S} p_i$. To solve this problem, we can first fix a monitoring structure π and solve for optimal incentives for any given monitoring team $S \in \pi$ as we did in Section 4.1. Proceeding as in that section yields the analog of Proposition 2: if b^*_S is an optimal scheme for monitoring team S, then there is a permutation $(i^*_{S1}, \ldots, i^*_{S|S|})$ of the agents in S such that $p_{i^*_{S1}} \geq \ldots \geq p_{i^*_{S|S|}}$ and, for each $j \in \{1, \ldots, |S|\}$,

$$b^*_{i^*_{Sj}} = \frac{p_{i^*_{Sj}}}{p_{i^*_{Sj}} - p_0} \frac{c}{f(S)} \prod_{j < j' \leq |S|} \frac{p_{i^*_{Sj'}}}{p_0}.$$

Hence, an optimal bonus continues to take the form of the partial-impementation bonus times a compensation for strategic risk, with the compensation factor for agent $i_{Sj} \in S$ now given by $\prod_{j < j' \leq |S|} \frac{p_{i^*_{Sj'}}}{p_0}$. Moreover, as in the case of agents who differ in their effort costs, an optimal permutation for monitoring team S places higher-skilled agents (i.e., agents with a higher probability of task completion under effort) earlier than lower-skilled agents, in order to provide assurance to the latter.[18]

Using the bonuses above to substitute in the principal's incentive cost, the principal's problem then reduces to choosing $\pi \in \Pi(\overline{n})$ to minimize

$$\sum_{S \in \pi, j \in \{1, \ldots, |S|\}} \frac{p_{i^*_{Sj}}}{p_{i^*_{Sj}} - p_0} \, c \prod_{j < j' \leq |S|} \frac{p_{i^*_{Sj'}}}{p_0},$$

where, for each $S \in \pi$ and each $j \in \{1, \ldots, |S|\}$, $i^*_{Sj} \in S$ and $p_{i^*_{S1}} \geq \ldots \geq p_{i^*_{S|S|}}$.

[18] This follows because given a permutation $(i_{S1}, \ldots, i_{S|S|})$, the bonus $b^*_{i_{Sj}}$ is submodular in $p_{i_{Sj}}$ and the compensation factor, the compensation factor is decreasing in j, and, for any $j \in \{1, \ldots, |S| - 1\}$, the compensation factor for agent i_{Sj} is lower the lower are the probabilities $p_{i_{Sj'}}$ for $j' > j$.

It is immediate that Proposition 3 continues to hold: every optimal monitoring structure $\pi^* = \{S_1^*, \ldots, S_{n^*}^*\}$ exhausts the principal's monitoring capacity by setting $n^* = \overline{n}$. However, without making further assumptions, it is not possible to derive a general characterization of the optimal partition of agents into $\overline{n}$ monitoring teams. The reason is that the compensation for strategic risk that an agent demands now depends not only on the size of his monitoring team and his location in the monitoring team's permutation, but also on the exact probabilities of task completion of other agents in the monitoring team. As such, the principal's optimal monitoring structure will also depend on the exact values of the agents' task completion probabilities.

Nevertheless, we can show that our results extend if we put more structure on the problem. For example, suppose that $p_i \in \{p_L, p_H\}$ for all $i \in M$ and some $p_0 < p_L < p_H \leq 1$. Then we obtain that every optimal monitoring structure is anti-assortative if p_L is sufficiently close to p_0. In this case, the number of agents of each type is the same across monitoring teams (up to integer constraints). Thus, as in our baseline model, the principal specifies monitoring teams that are heterogeneous within but homogeneous between them.

5.2 Beyond partitions

We have modeled a monitoring structure as a partition of the set of agents into monitoring teams. How would the principal's problem change if monitoring is not required to take the form of a partition? While a full analysis of this question is beyond the scope of our paper, we offer here some insights.

Consider first the class of deterministic monitoring structures. A non-partition structure would allow for overlapping monitoring teams, meaning that an agent could be assigned to multiple monitoring teams simultaneously. The applicability of such overlaps may vary depending on the context; for example, they would be infeasible if different monitoring teams must reside in different physical locations.

Even when feasible, overlapping monitoring teams may not be beneficial to the principal. Take the setting of Section 4.3 and suppose the principal assesses agents' performance by verifying the successful completion of tasks up to different points along the vector $(1, \ldots, m)$. For instance, in our hiring example in the Introduction, to specify a monitoring partition $\{\{1, 2\}, \{3, 4\}\}$, the principal verifies

the joint performance of tasks 1 and 2 (quality of interviewees) and the joint performance of all tasks (quality of hires) relative to those of 1 and 2. What would it mean to specify overlapping monitoring teams, such as $\{\{1,2,3\},\{3,4\}\}$? The principal would need to verify the joint performance of tasks 1, 2, and 3 (quality of offers) and the joint performance of all tasks relative to those of 1 and 2. However, in this case, the principal would also be able to specify the monitoring partition $\{\{1,2\},\{3\},\{4\}\}$, and such a partition always improves upon $\{\{1,2,3\},\{3,4\}\}$.

Once we move beyond partitions, in principle there is no reason to require that an agent's inclusion in a monitoring team be a binary decision. The principal might be able to specify fractional assignments of agents to monitoring teams, similar to those studied in Meyer (1994) and Chade and Eeckhout (2018) in the context of productive teams. Whether a fractional assignment is beneficial to the principal may depend on the assumptions that we make on the signal structure. In particular, a relevant modeling question is how the signal produced by a monitoring team would then be affected by the performance of agents who belong to the monitoring team only fractionally.

Finally, consider random monitoring structures. Absent constraints, if the principal can commit to any randomization over monitoring partitions, then agents' strategic risk can be eliminated at no cost. For example, for $\pi_i := \{\{i\}, M \setminus \{i\}\}$, suppose the principal randomizes over $\{\pi_i\}_{i\in M}$ and specifies a bonus for each agent $i \in M$ that is conditional on π_i being drawn and the agent delivering a good performance signal. Plainly, by effectively evaluating each agent's performance individually, such a mechanism would uniquely implement work at the partial-implementation cost.

There are a number of issues, however, with this approach. There is the usual concern that committing to a randomization may be difficult. More importantly, there is the problem that the random mechanism above respects the principal's capacity constraint ex post, namely after a partition is drawn, but not ex ante. For the randomization to be effective, we must thus assume either that the principal can set up the monitoring partition after agents have chosen effort, or that she can make the monitoring partition unobservable to the agents. Both of these are arguably strong assumptions. In reality, creating performance measures requires the principal to incur ex-ante costs. If agents can observe what performance signals are available to the principal, the capacity constraint applies ex ante, and

as a result randomizations cannot improve upon deterministic mechanisms.

5.3 Performance signals

Our model assumes that for each monitoring team that the principal specifies, she can only verify whether or not all agents in the monitoring team have successfully completed their tasks. That is, the principal can tell if something did not go well, but she cannot discern more than that. This signal structure corresponds to the "O-ring" production function in Kremer (1993) and in the benchmark setting of Winter (2004). Its main feature is that it is binary and supermodular.

While we have taken the signal structure as given, one could consider an augmented design problem in which the principal is able to choose the signal that she observes for each monitoring team, in addition to the monitoring partition and the bonus schemes. A detailed analysis of this augmented problem requires modeling the costs of different performance signals, which is beyond our scope. However, it is worth noting that the signal that we have assumed is in fact optimal in the augmented problem under simple conditions.

For instance, suppose that the principal is restricted to signals of joint performance that are binary and deterministic. That is, for each monitoring team, the principal can only choose what profiles of agents' task completion outcomes map to a good signal, with the remaining profiles mapping to a bad signal. Our model assumes that only one outcome profile maps to a good signal, namely that in which all agents in the monitoring team have completed their tasks. We can show that, within the proposed class, this signal is optimal if the individual probability of task completion under effort, p_1, is sufficiently close to 1.

To illustrate the idea, take a monitoring team and suppose that, unlike in our model, the principal observes a good signal whenever the number of agents who complete their tasks exceeds a strictly interior threshold. Then an agent has little incentive to work if he conjectures that sufficiently many other agents are working, as the marginal contribution of his effort to producing a good signal is then small. In fact, this marginal contribution is negligible if p_1 is high enough. As a consequence, in this case, the principal's cost of inducing all agents to work becomes arbitrarily high, regardless of whether she wishes to do so as some equilibrium outcome or as the unique one. Since the principal's incentive cost under

our signal structure remains bounded as p_1 increases, it follows that our signal structure is preferred for p_1 high enough.

References

Adhvaryu, Achyuta, Vittorio Bassi, Anant Nyshadham, and Jorge Tamayo, "No Line Left Behind: Assortative Matching Inside the Firm," 2020. Working paper.

Alchian, Armen A. and Harold Demsetz, "Production, Information Costs, and Economic Organization," *American Economic Review*, 1972, *62* (5), 777–795.

Baliga, Sandeep, "The Not-So-Secret-Agent: Professional Monitors, Hierarchies and Implementation," *Review of Economic Design*, 2002, *7*, 17–26.

Becker, Gary S., "A Theory of Marriage: Part I," *Journal of Political Economy*, 1973, *81* (4), 813–846.

Bernstein, Shai and Eyal Winter, "Contracting with Heterogeneous Externalities," *American Economic Journal: Microeconomics*, 2012, *4*, 50–76.

Chade, Hector and Jan Eeckhout, "Matching information," *Theoretical Economics*, 2018, *13*, 377–414.

Franco, April Mitchell, Matthew Mitchell, and Galina Vereshchagina, "Incentives and the Structure of Teams," *Journal of Economic Theory*, 2011, *146*, 2307–2332.

Gershkov, Alex and Eyal Winter, "Formal versus Informal Monitoring in Teams," *American Economic Journal: Microeconomics*, 2015, *7* (2), 27–44.

Halac, Marina, Elliot Lipnowski, and Daniel Rappoport, "Rank Uncertainty in Organizations," *American Economic Review*, 2021, *111* (3), 757–786.

_ , Ilan Kremer, and Eyal Winter, "Raising Capital from Heterogeneous Investors," *American Economic Review*, 2020, *110* (3), 889–921.

Holmström, Bengt, "Moral Hazard in Teams," *The Bell Journal of Economics*, 1982, *13* (2), 324–340.

Kambhampati, Ashwin and Carlos Segura-Rodriguez, "The Optimal Assortativity of Teams Inside the Firm," 2020. Working paper.

Kaya, Ayca and Galina Vereshchagina, "Partnerships versus Corporations: Moral Hazard, Sorting, and Ownership Structure," *American Economic Review*, 2014, *104* (1), 291–307.

Kremer, Michael, "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics*, 1993, *108* (3), 551–575.

McAfee, R. Preston and John McMillan, "Optimal Contracts for Teams," *International Economic Review*, 1991, *32* (3), 561–577.

Meyer, Margaret A., "The Dynamics of Learning with Team Production: Implications for Task Assignment," *Quarterly Journal of Economics*, 1994, *109* (4), 1157–1184.

Milgrom, Paul and John Roberts, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 1990, pp. 1255–1277.

Miller, David A. and Kareen Rozen, "Wasteful Sanctions, Underperformance, and Endogenous Supervision," *American Economic Journal: Microeconomics*, 2014, *6* (4), 326–361.

Miller, Nolan H., "Efficiency in Partnerships with Joint Monitoring," *Journal of Economic Theory*, 1997, *77* (2), 285–299.

Moldovanu, Benny, Aner Sela, and Xianwen Shi, "Contests for Status," *Journal of Political Economy*, 2007, *115* (2), 338–363.

Mookherjee, Dilip, "Optimal Incentive Schemes with Many Agents," *Review of Economic Studies*, 1984, *51* (3), 433–446.

_ **and Stefan Reichelstein**, "Dominant Strategy Implementation of Bayesian Incentive Compatible Allocation Rules," *Journal of Economic Theory*, 1992, *56* (2), 378–399.

Moriya, Fumitoshi and Takuro Yamashita, "Asymmetric-Information Allocation to Avoid Coordination Failure," *Journal of Economics & Management Strategy*, 2020, *29* (1), 173–186.

Picard, Pierre and Patrick Rey, "Incentives in Cooperative Research and Development," 1987. CEPREMAP Working Papers (Couverture Orange) 8739.

Prat, Andrea, "Should a Team Be Homogeneous?," *European Economic Review*, 2002, *46* (7), 1187–1207.

Rahman, David, "But Who Will Monitor the Monitor?," *American Economic Review*, 2012, *102* (6), 2767–2797.

Sákovics, József and Jakub Steiner, "Who Matters in Coordination Problems?," *American Economic Review*, 2012, *102* (7), 3439–3461.

Segal, Ilya R., "Contracting with Externalities," *Quarterly Journal of Economics*, 1999, *114*, 337–388.

_ , "Coordination and Discrimination in Contracting with Externalities: Divide and Conquer?," *Journal of Economic Theory*, 2003, *113*, 147–81.

Strausz, Roland, "Efficiency in Sequential Partnerships," *Journal of Economic Theory*, 1999, *85* (1), 140–156.

Winter, Eyal, "Incentives and Discrimination," *American Economic Review*, 2004, *94*, 764–773.

_ , "Transparency among Peers and Incentives," *RAND Journal of Economics*, 2010, *41* (3), 504–523.

A Appendix: Proofs

This Appendix provides formal proofs for our results. We obviate the proofs for Proposition 1 and Proposition 6 since they follow immediately from the arguments in the text.

A.1 Proof of Lemma 1

Fix a monitoring structure $\pi \in \Pi(\overline{n})$ and suppose b_S uniquely implements work in $S \in \pi$. Note first that there must be an agent $i_{S1} \in S$ who is willing to work under b_S when no other agent in S does. If this was not true, there would be an $\varepsilon > 0$ and a Nash equilibrium of the game induced by $b_S + \varepsilon$ in S in which no agent works. By supermodularity of the signal structure, it follows that i_{S1} is willing to work no matter what the other agents in S do.

We proceed by induction: take $j' \in \{2, \ldots, |S| - 1\}$ and suppose that for every $j \in \{2, \ldots, j'\}$, there is an agent i_{Sj} who is willing to work if agents $(i_{S1}, \ldots, i_{Sj-1})$ work, no matter what the other agents in S do. Then we claim there must be an agent $i_{Sj'+1}$ who is willing to work if agents $(i_{S1}, \ldots, i_{Sj'})$ work and the other agents in S do not. Otherwise, there would be an $\varepsilon > 0$ and a Nash equilibrium of the game induced by $b_S + \varepsilon$ in which agents $(i_{S1}, \ldots, i_{Sj'})$ work and the rest of the agents in S shirk. By supermodularity, it follows that $i_{Sj'+1}$ is willing to work if agents $(i_{S1}, \ldots, i_{Sj'})$ work, no matter what the other agents in S do.

A.2 Proof of Proposition 2

Fix a monitoring structure $\pi \in \Pi(\overline{n})$ and, without loss, a monitoring team $S \in \pi$. We proceed in two steps.

Step 1. Suppose that an incentive scheme b_S is optimal in monitoring team S. We show that there must be a permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S such that, for each $j \in \{1, \ldots, |S|\}$, agent i_{Sj} is indifferent between working and shirking when agents $(i_{S1}, \ldots, i_{Sj-1})$ work and agents $(i_{Sj+1}, \ldots, i_{S|S|})$ shirk.

By Lemma 1, there must be a permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S such that, for each $j \in \{1, \ldots, |S|\}$, agent i_{Sj} is willing to work if at least agents $(i_{S1}, \ldots, i_{Sj-1})$ work. By supermodularity, this holds if and only if, for each $j \in \{1, \ldots, |S|\}$, agent i_{Sj} is willing to work when agents $(i_{S1}, \ldots, i_{Sj-1})$ work and agents $(i_{Sj+1}, \ldots, i_{S|S|})$ shirk:

$$p_1^j p_0^{|S|-j} b_{i_{Sj}} - c_{i_{Sj}} \geq p_1^{j-1} p_0^{|S|-j+1} b_{i_{Sj}}. \tag{9}$$

We claim that optimality of the incentive scheme requires that, for some permu-

tation $(i_{S1}, \ldots, i_{S|S|})$, (9) hold with equality for each $j \in \{1, \ldots, |S|\}$. Suppose by contradiction that this is not the case; that is, suppose that there is an optimal incentive scheme with associated permutation $(i_{S1}, \ldots, i_{S|S|})$ and a bonus schedule such that (9) is (a weak inequality for all $j \in \{1, \ldots, |S|\}$ and) a strict inequality for some $j' \in \{1, \ldots, |S|\}$. Consider a perturbation in which we reduce $b_{i_{Sj'}}$ by $\delta > 0$ arbitrarily small while keeping all other bonuses unchanged. Since (9) was a strict inequality for j', this condition continues to be satisfied for all $j \in \{1, \ldots, |S|\}$. Moreover, the principal's incentive cost in (P) strictly decreases with the perturbation. It follows that the original scheme cannot be optimal.

Step 2. Suppose that an incentive scheme b_S is optimal in monitoring team S. We show that the implied permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S described in Step 1 must satisfy

$$c_{i_{S1}} \leq \ldots \leq c_{i_{S|S|}}. \tag{10}$$

By Step 1, an optimal incentive scheme implies a permutation $(i_{S1}, \ldots, i_{S|S|})$ of the agents in S such that, for each $j \in \{1, \ldots, |S|\}$,

$$p_1^j p_0^{|S|-j} b_{i_{Sj}} - c_{i_{Sj}} = p_1^{j-1} p_0^{|S|-j+1} b_{i_{Sj}},$$

or, equivalently,

$$b_{i_{Sj}} = \frac{p_1}{p_1 - p_0} \frac{c_{i_{Sj}}}{f(S)} r_{Sj},$$

where r_{Sj} is defined in equation (3). Hence, the principal's incentive cost for monitoring team S is given by

$$\sum_{j \in \{1, \ldots, |S|\}} \frac{p_1}{p_1 - p_0} c_{i_{Sj}} r_{Sj}. \tag{11}$$

Suppose by contradiction that the permutation $(i_{S1}, \ldots, i_{S|S|})$ does not satisfy (10). Then there exists $j' \in \{1, \ldots, |S| - 1\}$ such that $c_{i_{Sj'}} > c_{i_{Sj'+1}}$. Consider a perturbation that swaps agents $i_{Sj'}$ and $i_{Sj'+1}$ in the permutation. Note that this swap only affects the terms j' and $j' + 1$ of the sum in (11). The change in the principal's incentive cost for monitoring team S is thus equal to

$$\frac{p_1}{p_1 - p_0} (r_{Sj'+1} - r_{Sj'}) \left(c_{i_{Sj'}} - c_{i_{Sj'+1}} \right).$$

Since $r_{Sj'+1} < r_{Sj'}$ and $c_{i_{Sj'}} > c_{i_{Sj'+1}}$, this expression is strictly negative. It follows that the original scheme cannot be optimal.

A.3 Proof of Proposition 3

As shown in the text, the principal's incentive cost under a monitoring structure $\pi \in \Pi(\overline{n})$ is given by

$$\sum_{S \in \pi, j \in \{1,\ldots,|S|\}} \frac{p_1}{p_1 - p_0} \, c_{i^*_{Sj}} \, r_{Sj},$$

where, for each $S \in \pi$ and each $j \in \{1, \ldots, |S|\}$, $i^*_{Sj} \in S$ and $c_{i^*_{S1}} \leq \ldots \leq c_{i^*_{S|S|}}$.

To prove the first claim, suppose by contradiction that $\pi = \{S_1, \ldots, S_n\}$ is optimal with $n < \overline{n}$. Since $\overline{n} \leq m$, there is $S_k \in \pi$ with $|S_k| > 1$. Take $\iota \in S_k$ and define $j(\iota)$ by $i^*_{S_k j(\iota)} = \iota$. (That is, $j(\iota)$ is the location of agent ι in an optimal permutation for monitoring team S_k.) We construct a new monitoring structure, $\pi' = \{S'_1, \ldots, S'_{n'}\}$, which differs from π only in that S_k is split into two monitoring teams. Specifically, let $n' = n+1$; $S'_\ell = S_\ell$ for each $\ell \in \{1, \ldots, k-1\}$; $S'_k = S_k \backslash \{\iota\}$; $S'_{k+1} = \{\iota\}$; and $S'_\ell = S_{\ell-1}$ for each $\ell \in \{k+2, \ldots, n+1\}$. The change in the principal's incentive cost from using π' instead of π, divided by the constant $\frac{p_1}{p_1 - p_0}$, is equal to

$$\sum_{j \in \{1,\ldots,j(\iota)-1\}} c_{i^*_{S_k j}} (r_{S'_k j} - r_{S_k j}) + c_\iota \left(1 - r_{S_k j(\iota)}\right).$$

Observe that $r_{S_k j(\iota)} \geq 1$, strictly if $j(\iota) < |S_k|$, and $r_{S_k j} > r_{S'_k j}$ for each $j \in \{1, \ldots, j(\iota) - 1\}$. Hence, the monitoring structure π' yields a strictly lower incentive cost than π, implying that π cannot be optimal.

To prove the second claim, take a monitoring capacity $\overline{n}$ and an optimal monitoring structure $\pi = \{S_1, \ldots, S_n\} \in \Pi(\overline{n})$. By our first claim, $n = \overline{n}$. Now consider a monitoring capacity $\overline{n}' > \overline{n}$. We can perform a perturbation to π analogous to that described above to construct a new monitoring structure $\pi' = \{S_1, \ldots, S_{n'}\}$ with $n' = \overline{n} + 1$. By analogous reasoning as above, the principal's incentive cost under π' is strictly lower than that under π. Since $\pi' \in \Pi(\overline{n}')$, the claim follows.

A.4 Proof of Proposition 4

Suppose by contradiction that $\pi = \{S_1, \ldots, S_{\overline{n}}\}$ is optimal and $|S_k| - |S_\ell| > 1$ for some $S_k, S_\ell \in \pi$. Let $\iota := i^*_{S_k 1}$. We perform a perturbation in which we move agent ι from S_k to S_ℓ. Specifically, we construct $\pi' = \{S'_1, \ldots, S'_{\overline{n}}\}$ with $S'_t = S_t$ for all $t \neq k, \ell$; $S'_k = S_k \setminus \{\iota\}$; and $S'_\ell = S_\ell \cup \{\iota\}$. We show that the perturbation strictly lowers the principal's incentive cost when placing ι first in the permutation for monitoring team S'_ℓ (which implies that it also strictly lowers the principal's incentive cost when taking an optimal permutation for S'_ℓ). The change in such cost, divided by the constant $\frac{p_1}{p_1 - p_0}$, is equal to

$$c_\iota \left(r_{S'_\ell 1} - r_{S_k 1} \right),$$

or, equivalently,

$$c_\iota \left[\left(\frac{p_1}{p_0} \right)^{|S'_\ell| - 1} - \left(\frac{p_1}{p_0} \right)^{|S_k| - 1} \right].$$

Since $|S'_\ell| = |S_\ell| + 1$ and we have assumed $|S_\ell| + 1 < |S_k|$, this expression is strictly negative.

A.5 Proof of Proposition 5

Suppose that $\pi = \{S_1, \ldots, S_{\overline{n}}\}$ is optimal. If it is anti-assortative, we are done. Suppose instead that this is not true. By definition, for every $\phi \in \Phi$, there is some $S \in \pi$ containing two agents of the same rank. By Proposition 4 and our characterization of optimal incentives, it follows that there must be $S_k, S_\ell \in \pi$, $i^*_{S_k j} \in S_k$, and $i^*_{S_\ell j'} \in S_\ell$ such that $j < j'$ and $c_{i^*_{S_k j}} > c_{i^*_{S_\ell j'}}$. Take S_k and S_ℓ with the highest indices j and j' for which this is true, and let $\iota_k := i^*_{S_k j}$ and $\iota_\ell := i^*_{S_\ell j'}$. We perform a perturbation in which we assign ι_k to S_ℓ and ι_ℓ to S_k. Specifically, we construct $\pi' = \{S'_1, \ldots, S'_{\overline{n}}\}$ with $S'_t = S_t$ for all $t \neq k, \ell$; $S'_k = S_k \cup \{\iota_\ell\} \setminus \{\iota_k\}$; and $S'_\ell = S_\ell \cup \{\iota_k\} \setminus \{\iota_\ell\}$. We show that the perturbation weakly lowers the principal's incentive cost when placing ι_k in position j' in the permutation for S'_ℓ and ι_ℓ in position j in the permutation for S'_k (which implies that it also lowers the principal's incentive cost when taking optimal permutations for S'_ℓ and S'_k).

The change in such cost, divided by the constant $\frac{p_1}{p_1-p_0}$, is equal to

$$c_{\iota_k}\left(r_{S'_\ell j'} - r_{S_k j}\right) + c_{\iota_\ell}\left(r_{S'_k j} - r_{S_\ell j'}\right).$$

Since $|S'_k| = |S_k|$ and $|S'_\ell| = |S_\ell|$, this expression simplifies to

$$(c_{\iota_k} - c_{\iota_\ell})\left[\left(\frac{p_1}{p_0}\right)^{|S_\ell|-j'} - \left(\frac{p_1}{p_0}\right)^{|S_k|-j}\right].$$

By Proposition 4 and $j < j'$, we have $|S_k| - j \geq |S_\ell| - j'$. Since $c_{\iota_k} > c_{\iota_\ell}$, it follows that this expression is weakly negative.

Observe that if the perturbed monitoring structure is not anti-assortative, then we can perform the same perturbation again for the next highest indices $j < j'$ for which there are $S_k, S_\ell \in \pi$, $i^*_{S_k j} \in S_k$, and $i^*_{S_\ell j'} \in S_\ell$ with $j < j'$ and $c_{i^*_{S_k j}} > c_{i^*_{S_\ell j'}}$. This procedure yields a new monitoring structure $\pi' = \{S'_k, \ldots, S'_{\overline{n}}\}$ that is anti-assortative. Since the original monitoring structure was optimal and each perturbation weakly increases the principal's objective, it follows that this new monitoring structure is also optimal.

Remark 1. *If $m/\overline{n}$ is an integer, then it follows from this proof that every optimal monitoring structure must be anti-assortative. Specifically, in this case, the claim in Proposition 4 yields that every optimal monitoring structure must have all monitoring teams of equal size, and therefore the perturbations considered above yield a strict (rather than weak) decline in the principal's incentive cost.*

A.6 Proof of Proposition 7

Consider the constrained problem. As claimed in the text, the results in Lemma 1, Proposition 2, and Proposition 3 continue to apply. We prove each part of Proposition 7 in order.

Part 1: Let $c_i \in \{c_L, c_H\}$ for all $i \in M$. Suppose by contradiction that $\pi = \{S_1, \ldots, S_{\overline{n}}\}$ is optimal with $|S_k| - |S_\ell| > 1$ and $|\{i \in S_k : c_i = c_H\}| > |\{i \in S_\ell : c_i = c_H\}|$ for some $k, \ell \in \{1, \ldots, \overline{n}\}$ such that either $\ell = k+1$ or $\ell = k-1$. Without loss, assume $\ell = k+1$. Take $\iota \in S_k$ such that $\iota + 1 \in S_{k+1}$. We show that moving agent ι from the larger monitoring team S_k to the smaller

monitoring team S_{k+1} strictly lowers the principal's incentive cost. Construct a new monitoring structure, $\pi' = \{S'_1, \ldots, S'_{\overline{n}}\}$, which differs from π only in that ι is assigned to S_{k+1}. Specifically, let $S'_t = S_t$ for all $t \neq k, k+1$; $S'_k = S_k \setminus \{\iota\}$; and $S'_{k+1} = S_{k+1} \cup \{\iota\}$. Define $j(\iota)$ by $i^*_{S_k j(\iota)} = \iota$ and $j'(\iota)$ by $i^*_{S'_{k+1} j'(\iota)} = \iota$. (That is, $j(\iota)$ is the location of agent ι in an optimal permutation for monitoring team S_k, and $j'(\iota)$ is his location in an optimal permutation for monitoring team S'_{k+1}.) There are two cases to consider:

Case 1: $c_\iota = c_L$. In this case, we can without loss take $j(\iota) = j'(\iota) = 1$. The change in the principal's incentive cost from using π' instead of π, divided by the constant $\frac{p_1}{p_1 - p_0}$, is thus equal to

$$c_L \left(r_{S'_{k+1}1} - r_{S_k 1} \right),$$

or, equivalently,

$$c_L \left[\left(\frac{p_1}{p_0} \right)^{|S'_{k+1}|-1} - \left(\frac{p_1}{p_0} \right)^{|S_k|-1} \right].$$

Since $|S'_{k+1}| = |S_{k+1}| + 1$ and we have assumed $|S_{k+1}| + 1 < |S_k|$, this expression is strictly negative. Hence, the monitoring structure π' yields a strictly lower incentive cost than π, implying that π cannot be optimal.

Case 2: $c_\iota = c_H$. In this case, we can without loss take $j(\iota) = |S_k|$ and $j'(\iota) = |S'_{k+1}|$. Let $m_{Lk} := |\{i \in S_k : c_i = c_L\}|$ and $m_{Lk+1} := |\{i \in S_{k+1} : c_i = c_L\}|$. The change in the principal's incentive cost from using π' instead of π, divided by the constant $\frac{p_1}{p_1 - p_0}$, is thus equal to

$$\begin{aligned} &\sum_{j \in \{1, \ldots, m_{Lk}\}} c_L (r_{S'_k j} - r_{S_k j}) + \sum_{j \in \{m_{Lk}+1, \ldots, |S_k|-1\}} c_H (r_{S'_k j} - r_{S_k j}) \\ + &\sum_{j \in \{1, \ldots, m_{Lk+1}\}} c_L (r_{S'_{k+1} j} - r_{S_{k+1} j}) + \sum_{j \in \{m_{Lk+1}+1, \ldots, |S_{k+1}|\}} c_H (r_{S'_{k+1} j} - r_{S_{k+1} j}). \end{aligned}$$

Since $|S'_k| = |S_k| - 1$ and $|S'_{k+1}| = |S_{k+1}| + 1$, this expression simplifies to

$$\left(1 - \frac{p_1}{p_0}\right) \left\{ \begin{array}{c} \sum\limits_{j \in \{1,\ldots,m_{Lk}\}} c_L \left(\frac{p_1}{p_0}\right)^{|S_k|-1-j} + \sum\limits_{j \in \{m_{Lk}+1,\ldots,|S_k|-1\}} c_H \left(\frac{p_1}{p_0}\right)^{|S_k|-1-j} \\ - \sum\limits_{j \in \{1,\ldots,m_{Lk+1}\}} c_L \left(\frac{p_1}{p_0}\right)^{|S_{k+1}|-j} - \sum\limits_{j \in \{m_{Lk+1}+1,\ldots,|S_{k+1}|\}} c_H \left(\frac{p_1}{p_0}\right)^{|S_{k+1}|-j} \end{array} \right\}.$$

Since the contradiction assumptions imply $|S_k| - 1 > |S_{k+1}|$ and $|S_k| - 1 - m_{Lk} \geq |S_{k+1}| - m_{Lk+1}$, the expression above is strictly negative. Hence, the monitoring structure π' yields a strictly lower incentive cost than π, implying that π cannot be optimal.

Part 2: Let $c_i < c_{i+1}$ for each $i \in M$. Suppose by contradiction that $\pi = \{S_1, \ldots, S_{\overline{n}}\}$ is optimal with $|S_k| < |S_{k+1}|$ for some $k \in \{1, \ldots, n-1\}$. Take $\iota \in S_{k+1}$ such that $\iota - 1 \in S_k$. We perform a perturbation in which we assign ι to S_k. Specifically, we construct $\pi' = \{S'_k, \ldots, S'_{\overline{n}}\}$ with $S'_t = S_t$ for all $t \neq k, k+1$; $S'_k = S_k \cup \{\iota\}$; and $S'_{k+1} = S_{k+1} \setminus \{\iota\}$. Note that since $c_i < c_{i+1}$ for each $i \in M$, ι must be placed first in any optimal permutation for S_{k+1} and last in any optimal permutation for S'_k. To show that the perturbation strictly reduces the principal's incentive cost, it is thus sufficient to show that it weakly reduces the principal's incentive cost when ι is placed first in the permutations for both S_{k+1} and $S_{k'}$. The change in such cost, divided by the constant $\frac{p_1}{p_1 - p_0}$, is equal to

$$c_\iota \left(r_{S'_k 1} - r_{S_{k+1} 1}\right),$$

or, equivalently,

$$c_\iota \left[\left(\frac{p_1}{p_0}\right)^{|S'_k|-1} - \left(\frac{p_1}{p_0}\right)^{|S_{k+1}|-1}\right].$$

Since $|S'_k| = |S_k| + 1$ and we have assumed $|S_k| < |S_{k+1}|$, this expression is weakly negative.