American Economic Journal: Microeconomics 2014, 6(4): 326–361
http://dx.doi.org/10.1257/mic.6.4.326

Wasteful Sanctions, Underperformance, and Endogenous Supervision[†]

By David A. Miller and Kareen Rozen*

We study optimal contracting in team settings where agents have many opportunities to shirk, task-level monitoring is needed to provide useful incentives, and it is difficult to write individual performance into formal contracts. Incentives are provided informally, using wasteful sanctions like guilt and shame, or slowed promotion. These features give rise to optimal contracts with underperformance, forgiving sanctioning schemes, and endogenous supervision structures. Agents optimally take on more assigned tasks than they intend to complete, leading to the concentration of supervisory responsibility in the hands of one or two agents. (*JEL* D82, D86, J41, M12, M54)

Bob and Carol are partners on a consulting team. Suppose Bob and Carol are each assigned two of the firm's four clients. In an interaction with each client firm, there is some probability that the assigned consultant (say Bob) thinks of an innovative solution to help that firm improve its business model. If so, it is feasible for Bob to write a strong report for the client, convincingly proposing the solution and supporting it with quantitative analysis. If he chooses not to exert the required effort, or if a good idea does not occur to him in the first place, he can simply write a low-quality report. When Carol subsequently examines Bob's work and finds that his recommendation to a client is weak, she cannot tell whether Bob intentionally shirked. The only instruments for motivating Bob are wasteful sanctions that do not transfer utility to Carol. For instance, she can berate him in front of other coworkers, give him a poor evaluation, or even try to have him fired.

Suppose Carol monitors Bob's work for both of his clients. Taking incentives into account, is it socially optimal for him to perform to his full ability whenever doing a good job is feasible? Not if the feasibility of a task is uncertain. Then it may be best for Bob to set a *work target* of completing at most one high-quality report, even if two are feasible. Still, being assigned two tasks, rather than just one, provides Bob

*Miller: University of Michigan, Department of Economics, 611 Tappan St., Ann Arbor, MI 48109 (e-mail: econdm@umich.edu); Rozen: Yale, Dept. of Economics and the Cowles Foundation for Research in Economics, 30 Hillhouse Ave., New Haven, CT 06511 (e-mail: kareen.rozen@yale.edu). This paper was previously titled "Optimally empty promises and endogenous supervision." We are grateful to Larry Samuelson, Todd Sarver, Joel Sobel, Juuso Välimäki, Joel Watson, seminar participants at Northwestern University and Yale University, and several anonymous referees for helpful comments and suggestions. Max Dvorkin and Aniela Pietrasz provided excellent research assistance. Miller thanks Microsoft Research, Yale University, and the Cowles Foundation for Research in Economics for hospitality and financial support, the National Science Foundation (NSF) for financial support (award SES–1127643), and his former colleagues at UCSD for all their support. Rozen thanks the NSF for financial support (award SES–0919955).

[†]Go to http://dx.doi.org/10.1257/mic.6.4.326 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

a buffer of one extra task he could potentially exert effort on, in case the other task turns out to be infeasible. In the presence of wasteful sanctions, it may be best for Carol to simply accept Bob's underperformance, rather than force him to exert effort whenever feasible. In particular, Carol should not sanction him unless both his tasks fail her inspection; i.e., she should be forgiving. The likelihood of having good solutions for both of Bob's clients is relatively low, so the cost of his underperformance is relatively small, while forgiveness mitigates wasteful sanctions in the more likely event that only zero or one of his tasks are feasible.

We study a class of contractual settings that contains this example. There can be many agents on the team, each with limited capacity for performing and monitoring tasks. Each completed task yields a fixed benefit to the team but requires costly individual effort to complete, and some tasks are not feasible at all. Whether a task is feasible can be thought of as the outcome of a stochastic technological constraint (e.g., whether the problem is apparent or even has a solution, or whether the resources needed to pinpoint or solve it are available), but can also be thought of as intrinsic to the agent-task combination (e.g., whether the agent is aware of a solution, or whether the agent is distracted by exogenous factors).[1] When another agent monitors a task, she can tell whether it was completed, but if it was not completed she cannot tell whether it was feasible. A key innovation is that we allow the agents to trade off between performance and monitoring. For simplicity, we assume that an agent can use a unit of her capacity to either be assigned a task (i.e., find out if it is feasible and decide whether to complete it) or to monitor the task of another agent. For example, Carol may have to go through the entire case file to assess the quality of Bob's report, which is an opportunity cost on her time.

Our model presumes that it is not possible to commit to transfers that are contingent on inspection outcomes; instead any sanction imposed on an agent is pure waste. Since output is quantified at the level of the team, if at all, it is hard to write performance-based incentives at the individual level into a formal contract.[2] Even informal enforcement of inspection-contingent transfers is difficult. In a partnership, partners have limited liability, so the ability to enforce transfers of utility through changes in ownership shares may involve a sharp tradeoff with future incentives. In a firm, rewarding some team members for the failures of their peers harms morale, and can induce rent-seeking behavior.[3] Lacking the ability to enforce transfers, the agents must provide individual-level incentives informally, using wasteful instruments like guilt and shame (Kandel and Lazear 1992; Barron and Paulson Gjerde 1997; Carpenter et al. 2009; Knez and Simester 2001). When extreme sanctions

[1] We assume that a task that is infeasible for one agent cannot be reassigned to another agent. In Section IVE, we discuss why our qualitative results should hold even if tasks can be reassigned and reattempted.

[2] Incentives based on team-level output would merely amplify the social benefits of task completion that are already a primitive of our model. In Section V, we study a firm that hires a team of agents, offering them a formal contract that is linear in team output, in concert with an informal contract of wasteful sanctions. Because the formal contract cannot distinguish among the tasks completed by different agents, the firm optimally offers a team-level output bonus that is too small to solve the moral hazard problem.

[3] We hold that this is the case even if the teammate rewarded for the failure is not the one who reported it.

arise, typically the worst available is separation, with its attendant search and dislocation costs.[4]

Returning to our earlier example, notice that since monitoring and task assignments are fungible, Bob and Carol could allocate their responsibilities differently, while still assuring that four tasks are assigned to them and each task is monitored. For example, they could assign all four tasks to Bob, and Carol could use all her capacity to monitor him. Or they could assign one to Bob and three to Carol, with each monitoring all the tasks of the other. Are all these arrangements welfare-equivalent? No. Whenever it is optimal for one of them to underperform, it is strictly better to designate one partner as a "worker," who is assigned all the tasks, and the other as a "supervisor," who specializes in monitoring. Even though agents are identical and there are no returns to scale in the underlying production function, an endogenous supervisor emerges due to statistical complementarities that arise from optimal underperformance. If instead *full performance* (exerting effort on all feasible tasks) were optimal, it would make no difference how supervisory responsibility was allocated.

Our general results are organized as follows. First, we examine the structure of optimal contracts when there are enough supervisory resources to monitor every task. Section IIA considers the simple case of two identical agents with bounded capacity, one of whom is exogenously assigned to supervise the other. It is optimal for the worker to set a work target below capacity if λ —the independently and identically distributed probability that any given task is feasible—is neither too high nor too low. Section IIB addresses the question of who should supervise whom, while maintaining the restriction that every assigned task must be monitored. Whenever there is underperformance, it is strictly optimal for one agent to specialize in performing tasks while the other specializes in monitoring. Section III then shows how to economize on monitoring, by randomizing over which tasks to monitor. Now full performance can be implemented using only two units of capacity for monitoring, and is indeed optimal when λ is very high. Nonetheless, underperformance is still optimal for an intermediate range of λ. When underperformance is optimal, the allocation of monitoring responsibility impacts welfare. Even when it is not fully optimal, a single-supervisor arrangement attains strictly more than a $(N-1)/N$ fraction of the optimal welfare given any number of agents N. Section IV shows that our results are largely robust to various extensions of the model, and Section V shows that the same issues arise for teams within firms.

Our setting fits stylized characteristics of environments in which production occurs in teams and requires accumulated job-specific human capital. Professional partnerships often fit this bill, such as consulting and legal partnerships. If they lack natural measures of individual output, partners must monitor each other. For instance, an agent may face numerous tasks in a single workday, and yet output may be measurable only at the team level, as well as noisy or hard to quantify. In such

[4] Aoki (1998, 58) writes "unless the employee possesses special skills that might be needed elsewhere, the value of those skills, accumulated in the context of teamwork and internal personal networking, would by and large be lost." In essence, though their formal contracts are weak, team members earn an "efficiency wage," and are motivated by the prospect of losing it.

environments, agents face many opportunities to shirk, rather than one or several. Task-level monitoring is needed to provide useful incentives, and task-level monitoring is best accomplished by peers who are familiar with each others' tasks. Teams within firms face similar problems.[5]

This paper fits into the theory literature on partnerships and teams with moral hazard, but emphasizes a new perspective on teamwork by bringing together three features: peer monitoring, high-dimensional effort, and wasteful sanctions. These features provide the foundation for studying two important tradeoffs—between production and monitoring, and between performance and punishment—as well as the endogenous allocation of monitoring responsibility. Much of the literature on teams addresses contracts that depend on stochastic team output, and focuses on the problem of free-riding,[6] or allows for exogenously specified individual-level monitoring.[7] In contrast, our approach endogenizes individual-level monitoring by putting the agents in charge of monitoring each other.[8] We assume that assigning an agent to monitor her peers crowds out her own productivity.[9] This allows us to study the tradeoff between productive and supervisory activity at both the individual level and the team level, and to study the optimal assignment of agents into productive and supervisory roles.

Whereas the prior literature generally studies agents who exert effort along one dimension or several complementary dimensions,[10] in our model each task constitutes an independent dimension of effort.[11] This assumption imposes a natural structure on the stochastic relationship among effort, output, and monitoring, and enables us to make more specific predictions about task completion and supervision than would be possible with a single dimension of continuous effort.

[5] The inability to measure individual output often arises from complexity, which leads firms to endogenously organize their workers into teams. According to Lazear and Shaw (2007), from 1987 to 1996 "the percent of large firms with workers in self-managed work teams rose from 27 percent to 78 percent," and moreover "the firms that use teams the most are those that have complex problems to solve." Similarly, Boning, Ichniowski, and Shaw (2007) find that steel minimills with more complex production processes are more likely to organize workers into "problem-solving teams."

[6] For example: Legros and Matsushima (1991); Legros and Matthews (1993); d'Aspremont and Gérard-Varet (1998); Battaglini (2006); Coviello, Ichino, and Persico (2014). Free riding, of course, also arises in public goods problems (e.g., Palfrey and Rosenthal 1984).

[7] For example: Mirrlees (1997); Holmström (1982); Holmström and Milgrom (1991); McAfee and McMillan (1991); Miller (1997); Che and Yoo (2001); Laux (2001); Kvaløy and Olsen (2006); Carpenter et al. (2009); Matsushima, Miyazaki, and Yagi (2010).

[8] The costly state verification literature (Townsend 1979; Border and Sobel 1987; Williamson 1987; Mookherjee and Png 1989; Snyder 1999) also endogenizes monitoring probabilities, although not the allocation of monitoring responsibility. The debt-like contracts with low-powered incentives and random verification that arise in that literature bear some similarity to our optimal kinked-linear sanctioning schemes. For different reasons, low-powered incentives arise in both cases: there, because monitoring is costly; here because punishments are socially costly.

[9] Li and Zhang (2001) formalize Alchian and Demsetz (1972)'s conjecture that costly monitoring should be the responsibility of a residual claimant. Rahman (2012) and Rahman and Obara (2010) show the monitor need not be the residual claimant when a mediator can make correlated recommendations. Like Li and Zhang (2001), we show monitoring responsibilities are optimally given to one agent, but like Rahman (2012) and Rahman and Obara (2010), we need not give the monitoring agent residual claims.

[10] For example: Alchian and Demsetz (1972); Mirrlees (1976); Holmström (1982); McAfee and McMillan (1992); Kandel and Lazear (1992); Aoki (1994); Barron and Paulson Gjerde (1997); Che and Yoo (2001); Li and Zhang (2001); Battaglini (2006); Kvaløy and Olsen (2006); Carpenter et al. (2009).

[11] Matsushima, Miyazaki, and Yagi (2010) also study a model in which agents have private information about the feasibility of arbitrarily many independent tasks, but assume that monitoring is exogenous and utility is transferable. Holmström and Milgrom (1991), Legros and Matsushima (1991), Legros and Matthews (1993), Miller (1997), d'Aspremont and GérardVaret (1998), and Laux (2001) allow multi-dimensional effort, but their agents have no private information. Coviello, Ichino, and Persico (2014) study dynamic scheduling of many tasks under diseconomies of scope.

Finally, a majority of the literature assumes that all incentives are provided through monetary payments, such that only the imbalance must be burned (or given to a residual claimant).[12] Instead, we rule out formal monetary transfers, and focus on providing incentives through informal sanctions that are socially wasteful.[13] Heuristically, a framework with sanctions may be interpreted as the reduced-form of a repeated game with a sharply kinked Pareto frontier. Such sanctions are a natural instrument in an environment in which the agents cannot commit to inspection-contingent transfers. With wasteful sanctions, underperformance arises due to the tradeoff between performing all feasible tasks and providing the necessary incentives to do so.

I. Model and Preliminaries

Consider a team of $N \geq 2$ risk-neutral agents, each of whom may perform or monitor up to $M \geq 2$ tasks. There is a countably infinite set of tasks, each of which is an identical single-agent job. Any given task is feasible with independent probability $\lambda \in (0, 1)$. If a task is infeasible, then it cannot be completed. If a task is feasible, the agent performing it can choose whether to shirk or exert effort cost $c > 0$ to complete it. Shirking is costless, but yields no benefit to the team. If the agent exerts effort to complete the task, each member of the team (including him) receives an expected benefit b/N, where $b > c > b/N$. Hence, each task is socially beneficial, but no agent will exert effort without further incentives. To simplify exposition, we assume that monitoring requires zero effort cost. (Section IVC shows this can be relaxed without affecting our results.)

The timing of the game is as follows:

- At $\tau = 1$, each agent is assigned to perform some number of tasks.
- At $\tau = 2$, each agent privately observes the feasibility of each of his tasks, and, for each feasible task, privately decides whether to shirk or exert effort.
- At $\tau = 3$, agents monitor each other. An agent who was assigned p tasks at $\tau = 1$ can monitor up to $M - p$ of the other agents' tasks. Each task can be monitored by at most one agent, but the agents can employ an arbitrary correlation device to coordinate their monitoring activities. Conditional on being monitored, with probability 1, a completed task will pass inspection, and an uncompleted task will fail inspection.[14] The monitoring agent, however, cannot distinguish whether the task was infeasible or intentionally shirked.
- At $\tau = 4$, the agents reveal the results of their inspections.
- At $\tau = 5$, each agent can impose unbounded sanctions on other agents, at no cost to himself.

[12] For example: Alchian and Demsetz (1972); Mirrlees (1976); Holmström (1982); Holmström and Milgrom (1991); Legros and Matsushima (1991); McAfee and McMillan (1991); Legros and Matthews (1993); Miller (1997); d'Aspremont and GérardVaret (1998); Laux (2001); Battaglini (2006); Matsushima, Miyazaki, and Yagi (2010); Rahman and Obara (2010). Another literature sees bonuses and penalties as financially equivalent, but uses reference-dependent preferences to distinguish the incentive effects (e.g., Aron and Olivella 1994; Frederickson and Waller 2005).

[13] Wasteful sanctions are also studied by Kandel and Lazear (1992); Barron and Gjerde (1997); Che and Yoo (2001); and Carpenter et al. (2009) in different settings.

[14] Perfect monitoring simplifies the exposition; we discuss imperfect monitoring later.

We consider a setting in which it is not possible to commit to transfers that are contingent on inspection outcomes; instead, any sanction imposed on an agent is pure waste. We study perfect Bayesian equilibria of this game. Since the sanctions at $\tau = 5$ are unbounded and costless for each agent to impose, the agents can discourage any observable deviations from the equilibrium path—in particular, deviations at time $\tau = 1$ are immediately observable. Moreover, by the revelation principle it is without loss of generality to restrict attention to equilibria in which agents reveal their inspection results truthfully at time $\tau = 4$. Similarly, since monitoring is costless, we may ignore deviations from the equilibrium monitoring behavior at time $\tau = 3$. Accordingly, we limit attention to behavior along the equilibrium path. In such equilibria, the main concern is to discourage unobservable deviations at time $\tau = 2$. We call the specification of equilibrium-path behavior a *contract*, for reasons that we address in Remark 1 below. In what follows, for any countable set Z, $\Delta(Z)$, is the set of probability distributions over Z, and $\mathbb{N} \equiv \{0, 1, 2, \dots\}$.

DEFINITION 1: *A contract specifies, for each agent* $i = 1, 2, \dots, N$:

(*i*) *an assignment* $p_i \in \mathbb{N}$, *specifying how many tasks should be assigned to her*;

(*ii*) *a task completion strategy* $s_i : \mathbb{N} \to \Delta\mathbb{N}$, *with every realization less than or equal to the argument, specifying how many of her assigned tasks to complete among those feasible*;

(*iii*) *a monitoring distribution* $\rho_i : \mathbb{N}^2 \to [0, 1]$, *where* $\rho_i(f, a)$ *is the probability that f of agent i's tasks fail inspection when she completes a tasks; and*

(*iv*) *a sanctioning scheme* $v_i : \mathbb{N}^{2N} \to \mathbb{R}_-$, *specifying the net sanction imposed on her as a function of the numbers of tasks that pass and fail inspection across all players*.

A contract must respect each agent's bounded capacity. A contract is *feasible* if $p_i \leq M$ for all i; and there exists a correlated distribution over who should monitor which tasks, such that no agent j monitors herself or monitors more than $M - p_j$ tasks, and for each agent i, $\rho_i(\cdot, a)$ is the resulting distribution over how many tasks fail inspection when she performs a tasks. This allows for the possibility that not all assigned tasks are monitored, and that which tasks (and which agents) are monitored can be randomized; Section III considers these possibilities in detail. A contract is *incentive compatible* if no agent has an incentive to deviate from her task completion strategy, given the assignments, monitoring distribution, and sanctioning schemes. A contract is *optimal* if it maximizes the team's aggregate utility among feasible and incentive compatible contracts.

Remark 1 (Contractual interpetation): We refer to truthful equilibrium path behavior as a "contract" to emphasize that this game environment can also be interpreted as a contractual setting. Suppose some external principal offers the agents a contract in which the principal formally commits to pay each agent b/N

for each task completed by the team, and informally recommends assignments, task completion strategies, and sanctioning schemes. Then it should be a perfect Bayesian equilibrium for the agents to be obedient to the recommendations and report their inspection outcomes truthfully, as well as for the principal to implement the recommended sanctioning schemes. We investigate this principal-agent interpretation in Section V.

Remark 2 (Randomization, noncontingent transfers, and individual rationality): A more general space of contracts would allow the agents to employ randomized assignments. However, for our purposes it is without loss of generality to restrict attention to deterministic assignments. For any optimal contract employing random assignments, there would be an optimal deterministic assignment in the support of the randomization. If agents could opt out of the game before $\tau = 1$, then for any contract yielding positive social welfare the agents would be willing to accept the contract "behind the veil of ignorance," i.e., before their "roles" (as workers or supervisors) were randomly assigned. Alternatively, by using ex ante (noncontingent) transfers, it would be easy to spread the wealth so as to make everyone willing to accept the contract, no matter how asymmetric were the roles. In light of these possibilities, we do not impose individual rationality constraints on the contract.

Before formalizing the incentive compatibility constraints, we show that the relevant space of contracts can be simplified without loss of generality.

LEMMA 1: *There exists an optimal contract satisfying the following, for each agent i:*

(*i*) *The number of tasks agent i completes is a deterministic function of the number a of his tasks that are feasible* (*so, with some abuse of notation, let* $s_i(a)$ *be this number*);

(*ii*) *Agent i's sanction depends only on how many of his tasks failed inspection, so, without loss of generality, we rewrite the sanctioning scheme as* $v_i : \mathbb{N} \to \mathbb{R}_-$;

(*iii*) *"Upward" incentive compatibility constraints for task completion are slack: when* $a \leq p_i$ *tasks are feasible, agent i strictly prefers to complete* $s_i(a)$ *tasks over completing any number of tasks* $a' > s_i(a)$;

(*iv*) $s_i(s_i(a)) = s_i(a)$; *in addition,* $a \leq a'$ *implies* $s_i(a) \leq s_i(a')$.

An optimal contract chooses assignments $p = (p_i)_i$, task completion strategies $s = (s_i)_i$, monitoring distributions $\rho = (\rho_i)_i$, and sanctioning schemes $v = (v_i)_i$ to maximize

$$(1) \qquad \sum_{i=1}^{N} \sum_{a \leq p_i} \binom{p_i}{a} \lambda^a (1-\lambda)^{p_i - a} \Big(s_i(a)(b-c) + \sum_{f \leq p_i} v_i(f) \rho_i(f; s_i(a)) \Big),$$

subject to feasibility and downward incentive compatibility (IC)

$$(2)\quad \sum_{f\le p_i} v_i(f)\rho_i(f;\, s_i(a)) + s_i(a)\Big(\frac{b}{N} - c\Big) \;\ge\; \sum_{f\le p_i} v_i(f)\rho_i(f;\, a') + a'\Big(\frac{b}{N} - c\Big)$$

for each downward deviation $a' < s_i(a)$, each number of feasible tasks $a \le p_i$, and each agent i.

II. Underperformance and Endogenous Supervision

In this section, we show the optimality of underperformance and how it endogenously gives rise to optimal supervision structures. Throughout this section, we impose the restriction that every task must be monitored. A strategy s_i has *underperformance* if there is some number of tasks a such that $s_i(a) < a$. Otherwise (i.e., if $s_i(a) = a$ for all $a \le p_i$), the strategy involves *full performance*. Our results identify *work target strategies* as an important class of task completion strategies. A task completion strategy s_i is a work target strategy if there is a target p_i^* such that $s_i(a) = \min\{a, p_i^*\}$ for all $a \le p_i$. A work target strategy has underperformance if $p_i^* < p_i$; in that case, we say there is a *buffer* of $p_i - p_i^*$ tasks.

A. *A Worker and a Supervisor*

Before discussing how endogenous supervision may arise in Section IIB, we first examine the implications of a simple supervisory structure. Suppose the team consists of two members, and that the contract calls for one agent to be a "worker" who is assigned all the tasks, and the other to be a "supervisor" who monitors all the tasks. Because only one agent is completing tasks, we drop the i subscript and simply use p to denote the number of tasks that the worker is assigned, and s to denote his task completion strategy. The following theorem shows that underperformance arises in optimal worker-supervisor contracts.

THEOREM 1 (Worker-supervisor contracts): *Conditional on a worker-supervisor structure, there is an optimal contract such that:*

(*i*) *The worker is assigned M tasks, but uses a work target strategy, completing at most p^* feasible tasks. The supervisor monitors all M tasks.*

(*ii*) *No sanction is imposed on the worker up to a threshold of $M - p^*$ inspection failures, but each additional inspection failure results in a marginal sanction of $c - b/2$.*

(*iii*) *The work target satisfies $0 < p^* < M$ if $1 - \Big(2 - \frac{c}{b/2}\Big)^{1/M} < \lambda < \Big(\frac{c}{b/2} - 1\Big)^{1/M}$.*

(*iv*) *The work target p^** is *increasing in λ.*

Theorem 1 says that the optimal worker-supervisor contract has the worker complete only up to a work target of p^* tasks, even though the worker is assigned M tasks and the supervisor monitors each and every one. The cutoff p^*, which is increasing in the probability of task feasibility λ, is strictly positive whenever $1 - \left(2 - \frac{c}{b/2}\right)^{1/M} < \lambda$, and is strictly smaller than the number of tasks the worker was assigned whenever $\lambda < \left(\frac{c}{b/2} - 1\right)^{1/M}$. Recall that for $N = 2$, $\frac{b}{2} < c < b$, so $\frac{c}{b/2} \in (1, 2)$. Hence, the interval of λs for which the worker underperforms increases with both the capacity and cost-benefit ratio. Indeed, underperformance is guaranteed to be optimal if either M or $\frac{c}{b}$ is high enough.

COROLLARY 1: *Conditional on a worker-supervisor structure, for any* $\lambda \in (0, 1)$ *and* $\frac{c}{b}$ *there exists M sufficiently large that an optimal contract has underperformance; for any* $\lambda \in (0, 1)$ *and M there exists* $\frac{c}{b} < 1$ *sufficiently large that an optimal contract has underperformance.*

We prove Theorem 1 below. To understand the intuition for underperformance, note that even if the worker intends to perform all his tasks, some of them are likely to be infeasible because $\lambda < 1$, so he will incur sanctions anyway. Since sanctions are costly, it is possible to reduce the cost of sanctions by forgiving a few failures. However, the worker is able to move the support of the monitoring distribution. For example, if he is assigned ten tasks, and the threshold for being sanctioned is three failures, then he will never complete more than eight tasks, even if all ten are feasible. When λ is not too close to one, this tradeoff is resolved in favor of underperformance.

PROOF:

Suppose that the task completion strategy s is optimal and that the worker is optimally assigned p tasks. Note that if the worker completes $s(a)$ tasks when a tasks are feasible, then $p - s(a)$ of his tasks will fail inspection. Incentive compatibility of s requires that for all $a' < s(a)$,

$$v(p - s(a)) + s(a)\left(\frac{b}{N} - c\right) \geq v(p - a') + a'\left(\frac{b}{N} - c\right). \tag{3}$$

Let $p^* = \max_{a \leq p_i} s(a)$ be the largest number of tasks completed under strategy s (in light of Lemma 1, $p^* = s(p)$). Examination of equation (3) reveals that the expected sanction is minimized under the *kinked linear* sanctioning scheme $v(p - s(a)) = \left(\frac{b}{2} - c\right)\max\{(p - s(a)) - (p - p^*), 0\}$, which imposes no sanction when p^* or more tasks are completed, but a sanction of $\left(\frac{b}{2} - c\right)(p^* - s(a))$ whenever $s(a) < p^*$. This sanctioning scheme is kinked-linear in the number of tasks left uncompleted. Hence, s must have a work target of p^*. Thus far, point (ii) and part of point (i) are proven.

Substituting the kinked-linear sanctioning scheme into equation (1), the team's welfare reduces to

$$p^*\left(\frac{b}{2} - c\right) + \frac{b}{2}\sum_{a=0}^{p}\binom{p}{a}\lambda^a(1 - \lambda)^{p-a}\min\{a, p^*\}, \tag{4}$$

where the first term is the worker's payoff, and the second term is the positive externality he generates for the team. Note that equation (4) is maximized at $p = M$, proving the remaining part of point (i). By contrast, p^* has a positive effect on the second term but a negative effect in the first term, $\left(\text{since } \frac{b}{2} - c < 0\right)$. The second term, which we call the *truncated expectation*, has increasing differences in p^* and λ, leading to the monotone comparative statics in point (iv). Given that p^* is increasing in λ, there will be underperformance whenever (i) using $p^* = 1$ gives a larger value in equation (4) than does $p^* = 0$, to avoid the degenerate case in which the optimal number of assigned tasks may as well be zero; and (ii) using $p^* = M - 1$ gives a larger value in equation (4) than does $p^* = M$, so that the cutoff is strictly smaller than the number of tasks assigned. The interval in point (iii) then follows from algebra.

The optimal work target is closely related to the cost-benefit ratio of tasks. Picking the optimal work target in equation (4) requires increasing p^* until the point at which

$$(5)\quad p^*\left(\frac{b}{2} - c\right) + \frac{b}{2}\sum_{a=0}^{p}\binom{p}{a}\lambda^a(1-\lambda)^{p-a}\min\{a, p^*\}$$

$$\geq (p^* + 1)\left(\frac{b}{2} - c\right) + \frac{b}{2}\sum_{a=0}^{p}\binom{p}{a}\lambda^a(1-\lambda)^{p-a}\min\{a, p^* + 1\}.$$

Rearranging equation (5), the optimal work target is the smallest p^* satisfying

$$(6)\qquad \frac{c - \frac{b}{2}}{\frac{b}{2}} \geq \sum_{a=p^*+1}^{p}\binom{p}{a}\lambda^a(1-\lambda)^{p-a},$$

where the right-hand side of equation (6) is simply the probability that more than p^* tasks are feasible. Because $b > c > b/2$, the left-hand side of equation (6) is always between zero and one. The larger is the ratio $\frac{c}{b/2}$, the smaller is the optimal work target p^*. In particular, if the left-hand side is larger than one-half (which is the case when $\frac{c}{b/2} \geq \frac{3}{2}$), then, since we are looking for the *smallest* p^* satisfying equation (6), the optimal work target can be no more than the median number of feasible tasks (i.e., $M\lambda$, rounded either up or down). More generally, using a normal approximation to the binomial, $p^*/M = \lambda \pm O(M^{-3/2})$ for large M (see equation (A13) in the Mathematical Appendix).

B. *Endogenous Supervision*

Our findings that agents should optimally employ work target strategies, that those work targets are increasing in λ, and that the optimal sanctioning scheme is forgiving, are not specific to the worker-supervisor structure studied above. Indeed, the results of Theorem 1 extend to any contract with *complete monitoring*: for every task of every agent, there is another agent who monitors that task with probability one.

COROLLARY 2: *Conditional on complete monitoring, there is an optimal contract such that each agent i has a work target strategy, and the sanctioning scheme is kinked-linear. Each agent's work target p_i^* is increasing in λ.*

There are many possible complete monitoring contracts. In addition to a worker-supervisor contract, in which the agents are completely specialized, the agents could split the burdens of task performance and monitoring equally (if M is even), or in asymmetric proportions. For $N > 2$, yet more intricate possibilities exist. As the following result shows, in the presence of underperformance, these contracts are not payoff-equivalent.

THEOREM 2 (Complete monitoring): *Conditional on complete monitoring, a worker-supervisor contract is optimal when there are two agents. When $N > 2$ agents, the division of labor in an optimal complete monitoring contract includes at least one supervisor (who specializes in monitoring). Among the workers who are assigned tasks, if agent i is assigned more tasks than j ($p_i > p_j$), then he also has a higher work target than j ($p_i^* > p_j^*$). The optimality in this result is strict if and only if λ is such that there is underperformance under the optimal worker-supervisor contract.*

As seen in the proof below, whenever underperformance is optimal, supervision endogenously arises for statistical reasons, despite the symmetry of players and the independence of tasks. It is particularly intuitive to consider the case in which M and p^* are even, and compare a worker-supervisor contract in which the worker has a work target of $p^* < M$ to a symmetric contract in which both players are assigned $M/2$ tasks and have work targets of $p^*/2$. In both cases, the total numbers of assigned tasks and buffer tasks are the same. Suppose exactly p^* tasks turn out to be feasible. In the worker-supervisor contract all of them will be completed. However, in the symmetric contract all of them will be completed only if each player turns out to have exactly half of them. The fact that each player has a separate work target in the symmetric contract means there are two constraints to be satisfied, rather than just the one in the worker-supervisor contract. The same issue arises when comparing any two arbitrary contracts with the same total number of assigned tasks and the same total number of buffer tasks—whichever is the more asymmetric is superior.

PROOF:

Consider a complete monitoring contract in which each agent i is assigned $p_i \in \{0, 1, \ldots, M\}$ tasks and has a work target of p_i^*. Without loss of generality, under complete monitoring, suppose that $\sum_{i=1}^{N} p_i = \frac{NM}{2}$, and let $p_{\text{sum}}^* = \sum_{i=1}^{N} p_i^*$. In analogy to equation (4), if the sanctioning scheme is optimized then the team's welfare is given by

$$(7) \qquad \sum_{i=1}^{N}\left(p_i^*\left(\frac{b}{N} - c\right) + \frac{N-1}{N} b \sum_{a=0}^{p_i}\binom{p_i}{a}\lambda^a(1-\lambda)^{p_i-a}\min\{a, p_i^*\}\right).$$

Consider another complete monitoring contract (with a corresponding optimized sanctioning scheme) where each agent i is assigned $\tilde{p}_i$ tasks and has a work target of $\tilde{p}_i^*$, also having the property that $\sum_{i=1}^{N} \tilde{p}_i = \frac{NM}{2}$ and $p^*_{\text{sum}} = \sum_{i=1}^{N} \tilde{p}_i^*$. Since the sum of work targets is the same in each case, the welfare ranking of the two contracts is determined by their truncated expectations, $\sum_i \sum_{a=0}^{p_i} \binom{p_i}{a} \lambda^a (1-\lambda)^{p_i - a} \min\{a, p_i^*\}$, from equation (7). If the contracts featured full performance for each worker, the truncated expectations would be identical, so the allocation of monitoring responsibility would not matter. So instead suppose that the contracts feature underperformance: $\sum_i p_i^* < \sum_i p_i$. In this case, team welfare depends not just on p^*_{sum} but on how the work targets and assignments are allocated across workers. Indeed, note that each truncated expectation is supermodular in p and p^*, because the condition for increasing differences reduces to

$$(8) \qquad \sum_{a=p^*+1}^{p+1} \binom{p+1}{a} \lambda^a (1-\lambda)^{p+1-a} - \sum_{a=p^*+1}^{p} \binom{p}{a} \lambda^a (1-\lambda)^{p-a} > 0,$$

which holds because being assigned more tasks leads to a first-order stochastic improvement in the number of feasible tasks. Since the truncated expectation is zero when $p_i = p_i^* = 0$, supermodularity implies superadditivity. In particular, when $N = 2$, superadditivity immediately implies that the worker-supervisor contract dominates all others. For the case $N > 2$, note first there must be at least one pair of agents i, j for whom $p_i + p_j \leq M$, else the contract would violate complete monitoring. This again ensures that there must be an agent specializing in monitoring, because tasks could be reallocated between them for an improvement. Moreover, the rearrangement inequality of Lorentz (1953) for supermodular sums implies that to maximize equation (7), it must be that $p_i \geq p_j$ implies $p_i^* \geq p_j^*$.[15]

There are two different ways to interpret complete monitoring. Under one interpretation, units of capacity are not substitutable across performance and monitoring. Rather, there are $\frac{NM}{2}$ performance units and $\frac{NM}{2}$ monitoring units, and the problem is to allocate these units within the team. One reason for complete monitoring might be legal liability when tasks are left unmonitored. Even in limited liability partnerships, for instance, law firm partners can be held personally liable when another partner's work is not monitored (Fortney 1995; Richmond 2007–2008). Under a second interpretation, units of capacity are perfectly substitutable across performance and monitoring, and complete monitoring arises when the two activities receive equal allocations. In this second interpretation, complete monitoring is an ad hoc constraint. The next section relaxes this constraint, allowing the team to monitor less in order to accomplish more.

[15] The Lorentz-Fan rearrangement inequality says if $f: \mathbb{R}^k \to \mathbb{R}$ is a supermodular function, then $\sum_{i=1}^{n} f(x^i) \leq \sum_{i=1}^{n} f(x^{*i})$ for any collection of vectors $(x^1, \ldots, x^n)$, where x^{*i} is the "majorized" vector, which, for every dimension k, contains the ith largest component among $x_k^1, \ldots, x_k^n$. Here, $x^i = (p_i, p_i^*)$, and f transforms x^i into a summand in equation (7).

III. Trading Off Performance and Monitoring

In this section, we examine the optimality of underperformance when units of capacity are substitutable between performance and monitoring. For example, rather than monitor all the worker's tasks, the supervisor in the previous section could use some of her units of capacity toward performing tasks—in which case the worker would need to monitor some tasks as well. In particular, they may wish to allocate more units of capacity to performing tasks than to monitoring tasks (else, in view of Theorem 2, their original worker-supervisor arrangement was optimal). With less monitoring, however, they may not be able to implement finely tuned sanctioning schemes. These concerns raise several questions. How much capacity should they devote to monitoring? How much capacity should they devote to buffer tasks, and how much to tasks that they intend to perform? Does it still matter how monitoring responsibility is distributed?

We begin by applying some insights from the analysis in Section II, wherein each players' tasks were monitored with probability one. Suppose instead that of complete monitoring, we employ only M units of monitoring regardless of the number of players. For each player i with $p_i > 0$, there is an $\alpha_i \in (0, 1]$, such that with probability α_i all of i's tasks are monitored, and with probability $1 - \alpha_i$ none of her tasks are monitored. We refer to such monitoring as *probabilistically complete*. We claim that devoting more than a single agent's capacity toward monitoring reduces the capacity available for performing tasks without improving over complete monitoring with regard to incentives. Indeed, the same expected sanctions (and work target strategies) available under complete monitoring are also available under probabilistically complete monitoring, simply by scaling the sanctioning scheme appropriately. The amount of monitoring used by optimal contracts is thus bounded by the capacity of one agent, no matter how many agents are involved.[16]

LEMMA 2: *An optimal contract allocates at least* 2 *and at most M units of capacity to monitoring. Among those contracts allocating M units of capacity to monitoring*, *optimal monitoring is probabilistically complete*, *and the results of Corollary 2* (*kinked-linear sanctioning schemes*, *work target strategies*, *and monotonicity of* p_i *and* p_i^* *in* λ *for each agent* i) *carry over.*

PROOF:

Any incentives that can be generated when fewer than all of an agent's tasks are monitoring can also be generated under complete monitoring: whatever sanction the agent expects when completing exactly a tasks can be replicated when all the agent's tasks are monitored, simply by letting $v(p - a)$ equal that expected

[16] Depending on when capacity must be allocated, correlation effects may further favor having a single supervisor under this type of monitoring. Consider a team of three agents (Alice, Bob, and Carol), with $M = 4$. Suppose Alice is assigned four tasks, and Bob and Carol are each assigned two tasks and monitor two tasks. With probability $\frac{1}{2}$ Bob and Carol each monitor 2 (different) tasks Alice was assigned, and with probability $\frac{1}{2}$ Bob and Carol each monitor each other. But if Bob learns at $\tau = 1$ that he will monitor Alice, he will have no incentive to complete any tasks, since he will know that Carol will monitor Alice as well. Theorem 3 shows that a single supervisor optimally arises even if the realization of the monitoring arrangement is kept secret until $\tau = 3$.

sanction. Moreover, any incentives that can be generated under complete monitoring can be generated under probabilistically complete monitoring, by multiplying the sanctioning scheme v_i under complete monitoring by the factor $1/\alpha_i$, so that, taking expectations over whether she will be monitored, her *expected sanctioning scheme* is exactly v_i. Given that all possible incentive schemes are available under probabilistically complete monitoring, an optimal contract would not allocate more than M units of capacity toward monitoring when units of capacity are substitutable between performance and monitoring.

For the lower bound, observe that if zero units of capacity were devoted to monitoring then no tasks would be performed, so it would be equally good to allocate M units to monitoring. If one unit were devoted to monitoring, then the agent with that unit would not perform any tasks, so, again, it would be equally good to allocate M units to monitoring.

When seeking to characterize an optimal contract, the ability to trade off performance and monitoring generates some difficulties. If fewer tasks are being monitored than an agent was assigned, it will not generally be possible to compute the optimal sanctioning scheme and task completion strategies as we did in the previous section. Unlike the case of (probabilistically) complete monitoring, in which it is simple to make all relevant incentive constraints bind, the sanctioning scheme may have too few degrees of freedom relative to the task-completion strategy. Compounding this problem, there is a gap between ex post and expected sanctions: conditional on the tasks a player completed, inspection outcomes depend probabilistically on the monitoring distribution. Because sanctions are restricted to be negative, it may be impossible to generate the expected sanctioning scheme that makes a given combination of incentive constraints bind.[17]

Even without solving explicitly for the fully optimal contract, however, it is possible to bound the welfare loss from assigning all supervisory responsibility to one agent.

THEOREM 3: *For any λ, the best single-supervisor contract attains strictly more than a $(N-1)/N$ fraction of the welfare from an optimal contract. Moreover, among all contracts that allocate M units of capacity to monitoring, a single-supervisor contract is optimal when $N = 2$, and approximately optimal for M large enough when $N > 2$.*

PROOF:

The bound on the welfare loss relative to a fully optimal contract is calculated as follows. Given λ, the best contract among all single-supervisor contracts implements, for each of the $N-1$ other workers, a strategy s with some work target p^* using the kinked-linear sanctioning scheme given by $v(p - s(a))$

[17] In the online Appendix, we show that if there is an additional constraint that the contract must impose increasingly harsh marginal penalties, then the optimal contract is in fact kinked linear.

$= \left(\frac{b}{N} - c\right) \max\{p^* - s(a), 0\}$ for each a. In analogy to equation (4), the social welfare of this contract is simply $N - 1$ times

$$p^*\left(\frac{b}{N} - c\right) + \frac{N-1}{N} b \sum_{a=0}^{p} \binom{p}{a} \lambda^a (1 - \lambda)^{p-a} \min\{a, p^*\}. \tag{9}$$

By Lemma 2, equation (9) is the largest possible contribution to social welfare that an agent can have when all of his tasks are monitored. We claim that this is also the largest possible contribution an agent can have in general. Indeed, suppose that only $m < p$ of an agent's tasks are monitored, and that the optimal sanctioning scheme in that case yields a certain expected sanction when the agent completes exactly a tasks. As noted in the proof of Lemma 2, those incentives can be replicated when all the agent's tasks are monitored, simply by letting $v(p - a)$ equal that expected sanction. Therefore, a fully optimal contract cannot possibly achieve a welfare greater than N times equation (9), which could happen only in the counterfactual situation that all tasks could be monitored without actually allocating any capacity toward monitoring.

As in the previous section, it is best to assign all M monitoring units to one supervisor when $N = 2$; when $N > 2$, we use the normal approximation to the binomial distribution to show that having a single supervisor is approximately optimal (leaving open the conjecture that it is in fact exactly optimal among all contracts devoting M units of capacity to monitoring). The detailed argument is relegated to the Mathematical Appendix.

Both parts of Theorem 3 are illustrated in Figure 1 for the case $b = 2, c = 3/2$, $N = 3$, and $M = 4$. The lower solid curve depicts the highest social welfare under a single-supervisor monitoring scheme with probabilistically complete monitoring. This is indeed at least $(N - 1)/N = 2/3$ of the optimal social welfare, which is depicted by the upper solid curve; in fact, the two coincide for low λ. The lower solid curve, for the single-supervisor arrangement, also dominates the dotted curve, which depicts the highest welfare that is available under probabilistically complete monitoring using multiple supervisors. The two coincide for high λ ($\gtrapprox 0.9$), since the allocation of supervisory responsibility is irrelevant under full performance.

It is also easy to demonstrate Theorem 3 analytically in the simple case of N agents, each with a capacity of $M = 2$. (For larger M, the analysis is much more complex.) An optimal contract in this case must use exactly two monitoring units. These can either be allocated to one supervisor or distributed across two "minimal supervisors." Because $M = 2$, the two minimal supervisors are, in effect, completely monitoring each other in the sense of Section II. Hence, having two minimal supervisors is either strictly worse than replacing them with one worker and one supervisor (if full performance is not optimal), or equivalent to it (if full performance is optimal), because giving the two monitoring units to one agent does not leave any wasted capacity. Thus, the single-supervisor arrangement achieves the optimum welfare for any λ, even if full performance is optimal. To see when underperformance

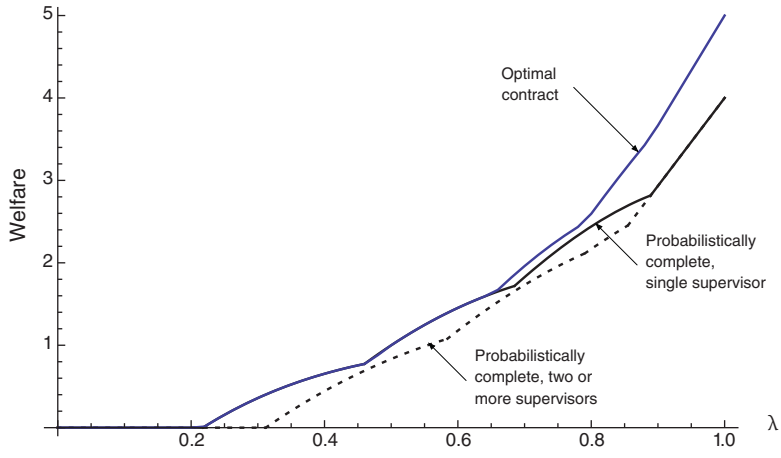

FIGURE 1. AN ILLUSTRATION OF THEOREM 3 IN THE CASE $b = 2$, $c = 3/2$, $N = 3$, and $M = 4$.

is optimal, observe that the welfare from a contract in which $N-1$ agents each have a target of only one task is $N-1$ times $\frac{b}{N} - c + 2b\frac{N-1}{N}\lambda - b\frac{N-1}{N}\lambda^2$, while full performance yields welfare equal to $N-1$ times $2\left(\frac{b}{N} - c\right) + 2b\frac{N-1}{N}$. Subtracting the latter from the former, the welfare difference is equal to $N-1$ times $c - \frac{b}{N} - b\frac{N-1}{N}\lambda^2$. Therefore underperformance (with a single-supervisor arrangement) dominates full performance if and only if $\lambda^2 \leq \frac{c}{b}\frac{N}{N-1} - \frac{1}{N-1}$. For $b = 2, c = 3/2$, and $N = 3$ (as in the previous example), this condition is approximately $\lambda < 0.79$.

More generally, we identify conditions under which underperformance is optimal by considering two particular types of supervisor schemes that generalize those in the analytical example above. First is the single-supervisor scheme with probabilistically complete monitoring, under which one player devotes all her capacity to supervising the rest. Second is a contract with *two minimal supervisors*, who each devote just one unit of capacity to monitoring both each other and the other players. In the latter case, the two monitoring units are used to generate a sanctioning scheme that implements a work-target strategy for the nonsupervisors. These two special contracts help us establish that our basic results on underperformance and endogenous supervision structures extend when monitoring may be incomplete and stochastic.

THEOREM 4: *For any capacity size M, number of agents N, and cost-benefit ratio $\frac{c}{b}$,*

(*i*) *If $\lambda < 1$ is sufficiently high, then the optimal contract is a full performance contract. An optimal full performance contract has exactly two "minimal supervisors" who each monitor one task.*

(*ii*) *There is $\lambda^* \in (0, 1)$, such that for all $\lambda \leq \lambda^*$, any full performance contract is dominated by a single-supervisor contract featuring probabilistically complete monitoring and underperformance.*

(*iii*) *Underperformance remains optimal for* λ *close to* 1 *if capacity and the cost-benefit ratio of tasks are moderately high: if* $\frac{c}{b} > \frac{2 + e/N}{2 + e} \approx 0.42 + O(N^{-1})$, *then a contract with two partial supervisors and underperformance dominates full performance in a neighborhood of* $\lambda = \frac{M - 2}{M - 1}$.

Theorem 4 is proved in the Mathematical Appendix. The first part of the theorem characterizes the best *full performance* contracts, and shows that there exist conditions under which full performance is optimal. The idea of the proof is as follows. Conditional on implementing full performance, monitoring can be minimized (and therefore the number of tasks completed can be maximized) by having two agents become minimal supervisors, who each monitor one task and have $M - 1$ tasks to perform. A correlated randomization device determines whether each supervisor monitors the other; with the remaining probability, the supervisors combine their monitoring capacity to monitor a worker randomly chosen from among the $N - 2$ other agents (each of whom is assigned M tasks). The sanctioning scheme is linear so that all agents are willing to complete all their feasible tasks, and all incentive constraints bind. Therefore, this is the optimal way to implement full performance. Full performance is clearly optimal when $\lambda = 1$, since in this case costly punishments are incurred with zero probability. We use a continuity argument to show full performance must also be optimal for λ in a neighborhood of 1. An implication of this part of the theorem is that whenever an optimal contract devotes more than two units of capacity to monitoring, then it must feature underperformance.

The third part of the theorem shows that even when λ is very close to 1 (that is, tasks are very likely to be feasible) underperformance is still optimal so long as M and $\frac{c}{b}$ are sufficiently large. In such cases, full performance is dominated by an underperformance contract with two minimal supervisors. Compared to a full performance contract, forgiving an agent's first failure yields a large benefit in terms of avoided punishment in the rather likely event that the agent has at least one infeasible task, at relatively low cost if it turns out that the agent's tasks are all feasible (because b is not too high compared to c).

Figure 2 illustrates the implications of this theorem for the case of $N = 4$, $M = 10$, and $\frac{c}{b} = \frac{1}{2}$, showing that full performance is dominated for a wide range of λ by the envelope of what can be attained using the two special classes of underperformance contracts described above. For $0.1 \lessapprox \lambda \lessapprox 0.6$, a single-supervisor contract with underperformance dominates full performance, because when λ is low, agents are likely to find many of their tasks to be infeasible. For $\lambda \gtrapprox 0.7$, tasks are too valuable to waste M units of capacity on monitoring (although more than $\frac{N - 1}{N} = 3/4$ of the maximal welfare could be achieved that way). However, even where it is optimal to allocate just two units of capacity to monitoring ($\lambda \gtrapprox 0.7$), underperformance still dominates full performance for $\lambda \lessapprox 0.8$.[18] Finally, Figure 2 also illustrates the maximum welfare attainable using work target

[18] Note that the cost-benefit ratio in this example does not satisfy the condition in part (ii) of Theorem 4. However, the condition is satisfied by the example in Figure 1, where $M = 4$. Indeed, underperformance is optimal for $\lambda \gtrapprox \frac{2}{3}$; this can be seen because the upper solid curve, which depicts the optimal social welfare in the example, is nonlinear in that region, whereas social welfare is linear under full performance.

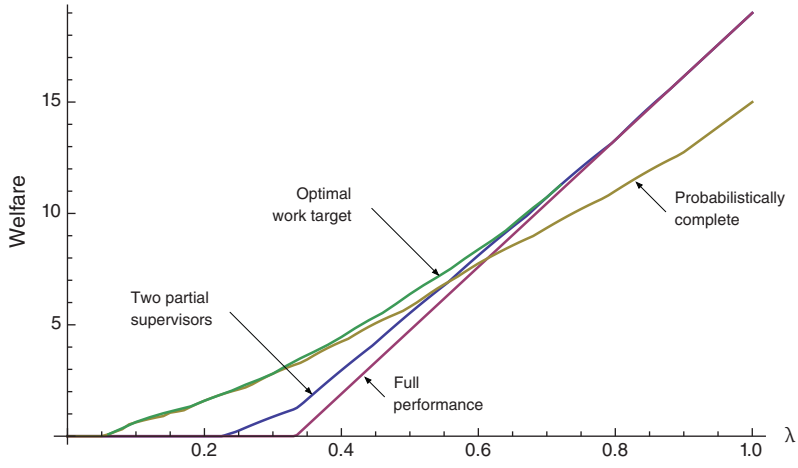

FIGURE 2. CONTRACTS WITH FOUR AGENTS, EACH WITH CAPACITY $M = 10$, TASK BENEFIT $b = 1$, AND TASK COST $c = 1/2$.

strategies. The gap between the optimal work target contract and the upper envelope of probabilistically complete and two-minimal-supervisor contracts is most noticeable for an intermediate range of λ. In this range, the optimal work target contract allocates more than 2 but less than M units of capacity to monitoring.

IV. Extensions

A. *Bounded Sanctions*

Although we have assumed that sanctions are unbounded, there could be some limit on the punishments that are available. For example, perhaps the only way to punish an agent is to fire him. In this case, gradations in the sanctioning scheme can still be achieved by using different probabilities of firing the agent for different inspection outcomes. So long as the disutility of being fired (with probability one) is sufficiently large to implement the harshest sanction needed to implement an optimal contract, our results would not change. Whether or not this constraint binds depends on the cost-benefit ratio of tasks, as well as the number of agents and their capacity for monitoring and performing tasks. Since the benefit of underperformance is to reduce the incidence of wasteful sanctions, our results on underperformance would be strengthened if sanctions were bounded. For instance, in the case of a worker-supervisor arrangement, the harshest possible sanction is needed to enforce full performance when all the worker's tasks are infeasible. If this harshest sanction is beyond a legal bound, then full performance simply cannot be implemented.

Given that the phenomenon of underperformance remains when sanctions are constrained, the way in which monitoring is allocated across agents still affects welfare. That is, it is still better to concentrate supervisory responsibility rather than disperse it. What may be affected by limits on sanctions is the feasibility of using a small number of supervisors. For example, if there are only one or two supervisors on a large team, then each agent is only monitored occasionally, so ex post

sanctions must be harsh to enforce good performance. If sanctions are constrained, agents must be monitored more often in order to preserve their incentives, so more resources must be devoted to monitoring (up to complete monitoring, if needed). So in the presence of constrained sanctions, in principle one could compute an optimal ratio of monitoring to performance, or of supervisors to workers.

B. *More Efficient Monitoring*

In our model, monitoring a task uses up the same amount of capacity as undertaking a task. In some settings, though, monitoring may be less resource-consuming than performing tasks. To fix ideas, suppose there are three or more players, and suddenly monitoring becomes twice as efficient, requiring only half a unit of capacity to monitor one task. Suppose also that it was optimal before to have two minimal supervisors (typically for high λ, as in Figure 2). Each such supervisor now has an unused, half-unit of capacity that can be used for additional monitoring. Since the optimal way to implement full performance uses only two monitoring units and a linear expected sanctioning scheme, additional monitoring units cannot improve the sanctions for full performance. However, additional monitoring units can lead to better sanctions for contracts featuring underperformance. Because monitoring is now cheaper, this can shift the tradeoff to further favor underperformance.

When underperformance is optimal, the allocation of monitoring responsibility affects social welfare. However, if probabilistically complete monitoring with a single supervisor was optimal before monitoring became cheaper (typically for moderately low λ), it may or may not remain optimal afterward, since the supervisor will have unused (and unmonitored) capacity. As long as λ is not too low to make use of that capacity, it may be preferable to also assign some monitoring responsibilities to a second (partial) supervisor, to make use of that unused capacity without suboptimally dispersing monitoring responsibilities into the hands of too many agents. Of course, aside from the two partial supervisors, none of the other agents' contracts are affected.

C. *Costly Monitoring*

The analysis thus far assumed that monitoring is costless, and therefore agents are indifferent over whether to monitor each other. If, however, monitoring requires nonverifiable, costly effort, the question of "who monitors the monitor" arises. Rahman (2012) shows that to provide incentives for monitoring, agents should occasionally shirk just to "test" the monitor. Since our model already generates optimal shirking (in the form of underperformance), we set monitoring costs to zero to highlight the fact that shirking arises from an entirely different mechanism. Adapting Rahman's argument, as follows, shows that the contracts we construct are robust to monitoring costs, without requiring any additional shirking.

Suppose that monitoring is costly. A monitor can always claim that a task passed inspection, but must exert effort to show that a task failed his inspection. To induce him to exert monitoring effort, the team can add an additional stage, $\tau = 6$, to their interaction. After the sanctions for failed tasks are implemented in $\tau = 5$, in

$\tau = 6$ each agent reports which tasks he himself completed. Agents are not punished for these reports, and are therefore willing to report truthfully. Whenever an agent reveals an uncompleted task in $\tau = 6$ that was not reported as failing inspection in $\tau = 4$, whichever teammate (if any) was supposed to monitor that task is punished. Because task feasibility is random, even under full performance there is positive probability that some tasks were not completed. Therefore a sufficiently large sanction induces faithful monitoring, and need not be incurred in equilibrium.

D. *Messages That Economize on Monitoring*

A recent literature studies the benefits of messages in contract design under private information.[19] In our model, incorporating messages can reduce the amount of monitoring needed. Consider the case in which all the tasks an agent is assigned are monitored. Matsushima, Miyazaki, and Yagi (2010) suggest that the principal should require an agent with private information to work on a certain number of tasks, which the agent should announce to the principal. Adapting this idea to our setting, we find that the same task completion strategies studied in earlier sections can be implemented using fewer than M units of capacity for monitoring. To see this, suppose an agent is assigned p tasks and his task completion strategy is s. Modify the contract to allow the agent to tell the other agents which p^* of his tasks to monitor, where $p^* = s(p)$ is the largest number of tasks he would ever complete. Clearly, the agent will include in his report all the tasks he has completed. Thus, no more than p^* tasks need to be monitored. Note that even if there are monitoring costs, the method in Section IVC for "monitoring the monitor" remains feasible since there is positive probability that fewer than p^* tasks were feasible.

Monitoring with messages allows more tasks to be assigned while still maintaining the same work target and expected sanctioning scheme. The optimal work target balances the resulting tradeoff between reducing the amount of monitoring and increasing the number of tasks completed. However, because the opportunity cost of underperformance is reduced for any given λ, full performance becomes even less attractive than before. Once again, different allocations of supervisory responsibility will not be welfare-equivalent under underperformance. With messages, a supervisor would have unused units of capacity, which could be allocated toward completing tasks. Since another agent must monitor him, optimal supervision structures with two partial supervisors in a team of N agents are again likely to arise.

E. *Reallocating Tasks*

We have assumed that if an agent is assigned a task that ends up being infeasible, she cannot reallocate that unit of capacity toward monitoring. Recall that the optimal way to implement full performance in our original setting is via a linear contract that makes all incentive constraints bind. As in the first part of Theorem 4, for a linear contract the same schedule of expected sanctions can be

[19] For example: Jackson and Sonnenschein (2007); Chakraborty and Harbaugh (2007); Matsushima, Miyazaki, and Yagi (2010); and Frankel (2014).

maintained by scaling the actual sanctions inversely with the number of monitoring slots employed. Consequently, the opportunity to allocate additional capacity toward monitoring is useful for reducing expected sanctions only if underperformance is optimal.

In our model, a task is either feasible or infeasible, regardless of whom it is assigned to. But there could be interesting team settings where the feasibility of any given task is idiosyncratic to each agent. This raises the possibility that agents could exchange tasks, to see whether the tasks that are infeasible for one might be feasible for another. Allowing for this would significantly alter the model, since it would require inserting both a task-trading phase and an additional task performance phase in between $\tau = 2$ and $\tau = 3$. At an intuitive level, however, allowing task trading would simply change the distribution over how many feasible tasks an agent might find, and an agent would still face a positive probability of finding fewer than he is willing to perform. Moreover, task trading introduces the new incentive constraint that an agent should not want to trade away a feasible task that he is supposed to fulfill. So there is still the same incentive problem of motivating him to perform tasks rather than claim they are infeasible for him. Qualitatively, the same kinds of results on underperformance and concentrated supervision would arise.

F. *Imperfect Monitoring*

We have assumed that when a shirked task is monitored, it will fail inspection with probability one. What if a shirked task that is monitored fails inspection with probability $\gamma \in (0, 1)$? The characterization of the optimal contract, conditional on M monitoring units, continues to hold for γ not too far from one; the other parts of Theorem 4 (points (i) and (iii)) continue to hold independently of γ.

The optimality of underperformance in point (iii) of Theorem 4 is unaffected by imperfect monitoring because the expected social welfare of the contract shown to dominate promise keeping is independent of γ. This contract is such that the sanctioning scheme depends on the number of failed inspections, and punishes only when the maximal number of failures is found. Letting F denote the number of tasks of a player that are monitored under this contract, the probability that F failures are found when completing a out of the p assigned tasks is given by $\gamma^F \binom{p-a}{F} / \binom{p}{F}$. The expected sanction conditional on completing a tasks is then given by $v(F)\gamma^F \binom{p-a}{F} / \binom{p}{F}$, which can be made independent of γ by scaling the sanction $v(F)$ by the factor $1/\gamma^F$.

Our characterization of the optimal contract conditional on M units of monitoring relies on being able to find a sanctioning scheme under which the expected sanctioning scheme makes all relevant incentive constraints bind. When the monitoring technology is sufficiently imperfect, such a schedule may not exist. Consider a simple case in which an agent is assigned 3 tasks, all of which are monitored. An uncompleted task fails inspection with probability $\frac{1}{4}$. Suppose for simplicity that $\frac{b}{N} - c = 1$. To induce a work target of $p^* = 2$, the optimal sanction would be 0 whenever at least 2 tasks are completed, -1 if 1 task is completed, and -2 if no

task is completed. Accounting for the probability of failing inspection, the sanctioning scheme $(v(0), v(1), v(2), v(3))$ should satisfy

$$\begin{pmatrix} \frac{27}{64} & \frac{27}{64} & \frac{9}{64} & \frac{1}{64} \\ \frac{9}{16} & \frac{3}{8} & \frac{1}{16} & 0 \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v(0) \\ v(1) \\ v(2) \\ v(3) \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \tag{10}$$

where the ath row and the fth column of the 4×4 matrix corresponds to the probability that f failures will be found when a tasks are completed. The unique solution to this system sets $v(0) = v(1) = 0$, $v(2) = -16$, and $v(3) = 16$. That is, the agent should receive a *reward* of 16 if the maximal number of failures are found! The difficulty here is that described by Farkas' Lemma: there is not always a negative solution to a linear system. By continuity, however, since a solution exists when the technology is perfect, one exists when the technology is not too imperfect. Indeed, in the example above, one can find a sanctioning scheme that generates the desired expected sanctioning scheme for any γ larger than about $\frac{1}{3}$.

V. Teams within Firms

When a team is embedded within a larger firm, the agents on the team may not directly benefit from the tasks they complete. The firm may be able to offer contractual bonuses that depend on the team's performance, but not on the performance of individual team members. At the level of individual performance, only wasteful sanctions (peer pressure, separation, etc.) are available. In this section, we show that the firm's problem of designing an optimal contract in this environment is similar to the problem the agents would face if they were partners, as characterized in the previous sections.

The firm hires a team of n agents to perform tasks and monitor each other. The firm reaps the entire benefit B from each task, but cannot observe who performed it. Agents have limited liability in terms of money, but can suffer from wasteful sanctions. For comparison to earlier results, we assume that $c < B < Nc$. The firm makes the agents a take-it-or-leave-it offer comprising:

- a fixed ex ante payment $t_i \geq 0$ for each i;
- a bonus $b \geq 0$, paid to the team for each completed task, and split equally among the agents so that each receives b/N;
- a randomization over *contracts* (assignments, task completion strategies, and sanctioning schemes; see Definition 1).

Each agent accepts the offer if and only if her expected utility from the offer is at least as high as her exogenous outside option, which is normalized to zero. Each agent's incentive compatibility constraint is simply equation (2). As discussed in

Remark 2, even if the firm's offer puts some agents into supervisory roles and others into productive roles, by randomizing the agents' roles after they accept the contract it can satisfy their individual rationality constraints as long as their expected utilities sum to at least zero. The firm's objective is to maximize

$$\sum_{i=1}^{N}\left(-t_i+\sum_{a=0}^{p_i}\binom{p_i}{a}\lambda^a(1-\lambda)^{p_i-a}\,s_i(a)(B-b)\right). \tag{11}$$

By straightforward elaboration on the usual argument, IR must bind in an optimal contract.[20] Substituting the binding IR constraints into equation (11) reduces the firm's objective function to the team's objective function (equation (1)), but with B in place of b. That is, when the firm hires the agents as a team, for any fixed bonus b its optimal contract is exactly the same as the optimal contract for the agents if they were partners. However, now the bonus is also a choice variable.

A bonus based on team output, naturally, is a crude instrument for providing incentives, since each team member receives b/N whenever any team member completes a task. Since profitability for the firm requires $b < B$, and yet $B/N < c$, the bonus alone cannot motivate the agents to perform and still yield positive profit for the firm. Hence, if the firm's optimal contract is nondegenerate, it must employ both a nonzero bonus and a nondegenerate sanctioning scheme. Moreover, observe that as the team gets larger it becomes more and more expensive to use the bonus for motivation. Indeed, since b is bounded above by B regardless of N, the bonus loses its motivational power in the limit as $N \to \infty$. At this limit, only sanctions provide incentives, so the bonus might as well be replaced by a fixed ex ante payment, since its only purpose is to meet the agents' IR constraints. As for the form of the firm's optimal contract, our previous conclusions still hold—underperformance and endogenous supervision arise for an intermediate range of λ.

These characteristics are consistent with stylized facts identified by Baker, Jensen, and Murphy (1987)—individual financial incentives are rare—and Oyer and Schaefer (2005)—broad-based group incentives are common. According to the model, these contractual features are optimal when the firm cannot formally monitor employees at the individual level, and must supplement its formal incentives at the team level with peer monitoring and informal sanctions at the individual level; i.e., industries where production is complex and requires accumulated job-specific human capital, as discussed in the introduction. For a striking example, Knez and Simester (2001) show that introducing a firm-level bonus scheme, complemented by peer monitoring and informal sanctions, increased on-time performance at Continental Airlines in the mid-1990s. The firm-wide bonus, coupled with the highly interdependent nature of on-time performance, provided each workgroup

[20] Consider a non-degenerate contract (in which agents complete some tasks) for which limited liability binds the bonus ($b = 0$) and IR is slack. The ex ante payments must be greater than zero (otherwise only a degenerate contract would be individually rational). But the firm can benefit from reducing the ex ante payments to zero, making it up to the agents by increasing the bonus to compensate. (When ex ante payments are zero, IR implies that the bonus satisfies $b \geq c$.) But since an increase in the bonus strengthens the agents' incentives, the firm can induce the same task performance at lower expected cost. Therefore, ex ante payments must be zero and the bonus must be nonzero. Further, IR cannot be slack, since the firm could impose marginally harsher sanctions to marginally reduce the bonus.

sufficient incentives to collectively prefer a high-effort equilibrium. At the individual level, high effort was supported by informal sanctions, where members of each workgroup would "monitor and sanction their colleagues to enforce the group decision" (Knez and Simester 2001, 746).

VI. Discussion

We study a model of teams in which agents optimally underperform relative to their abilities, and are "forgiven" for having done so. Underperformance buffers against the potential infeasibility of tasks, thereby minimizing costly sanctions. Underperformance arises even though buffering uses up capacity that could otherwise be allocated toward more detailed monitoring and finer, more attenuated sanctioning schemes.

Our model endogenously gives rise to optimal supervisory structures, despite the fact that all agents and tasks are identical. Although there is no inherent complementarity in task completion, increasing returns to a worker's task load when he underperforms arise from statistical complementarities: doubling both the number of tasks he is assigned and his work target of how many to complete more than doubles his social contribution. Consequently, it is best to have one agent be the "worker" and the other agent be the "supervisor," rather than have mixed roles. Under the assumption of unbounded liability, this intuition implies that there should be at most two supervisors, no matter how large the team. More realistically, a bound on liability would yield a lower bound on the ratio of supervisors to workers.

Introducing asymmetries into the model, even with complete information, may lead to additional interesting predictions. Suppose, for example, that the probability of task feasibility λ is player-specific. Then the least capable player should be performing as few tasks as possible, and using his resources toward supervision instead. This accords with the "Dilbert principle," which suggests that less productive team members should become supervisors (Adams 1996). Of course, if an agent who is better at performing tasks can also teach other agents, and if supervising and teaching are complementary, then it might instead be optimal for the most productive team members to supervise despite the loss of their production.

While the capacity constraints in our model serve the technical purpose of ensuring an optimal solution, they are also amenable to a bounded rationality interpretation. Although it is commonly assumed in contract theory that an agent's memory has unlimited capacity and perfect recall, evidence from psychology shows that working memory is both sharply bounded and imperfect.[21] One interpretation for the limiting resource is a bound on the number of tasks an agent can remember. A task in this

[21] A seminal paper by Miller (1956) suggests working memory capacity is about 7 ± 2 "chunks." A chunk is a set of strongly associated information—e.g., information about a task. More recently, Cowan (2000) suggests a grimmer view of 4 ± 1 chunks for more complex chunks. The economic literature studying imperfect memory includes Dow (1991); Piccione and Rubinstein (1997); Hirshleifer and Welch (2004); Bénabou and Tirole (2002); Wilson (2014); Kocer (2012). Mullainathan (2002) and Bodoh-Creed (2013) study updating based on data from long-term memory. There is also a literature on repeated games with finite automata, which can be interpreted in terms of memory constraints (e.g., Piccione and Rubinstein 1993; Cole and Kocherlakota 2005; Compte and Postlewaite 2008; Romero 2011), as well as work on self-delusion in groups (e.g., Bénabou 2013).

view contains detailed information, such as a decision tree, that is necessary to complete it properly.[22] Imperfect task feasibility may arise from being unable to remember all the necessary details for proper task completion. When tasks are complex, it may be impossible to fully specify their details in a convenient written form, such as a contract. As noted by Aoki (1988, 15), "the experience-based knowledge shared by a team of workers on the shopfloor may be tacit and not readily transferable in the form of formal language." Without a convenient way to fully specify a task, an agent who is assigned the task must expend memory resources to store the relevant details. Moreover, another agent may need to expend resources to store those details in order to be able to monitor him, leading to a tradeoff between performance and monitoring as in Section III. Coping with multiple complex tasks "may require more versatile workers' skills (deeper and broader information-processing capacities), which have not been considered essential in traditional hierarchies" (Aoki 1988, 31).

Mathematical Appendix

PROOF OF LEMMA 1:

Part (i): Suppose there is an optimal contract in which the assignment scheme is not deterministic. However, since the assignment scheme is realized publicly, there is an equally good contract that assigns probability 1 to whichever realization yields the highest welfare.

Part (ii): First, for agents to reveal their inspection results truthfully, their sanction must not depend on their announcements at time $\tau = 4$. Conditional on the monitoring scheme, an agent has no influence over whether other agents' tasks pass or fail inspection. So for any sanctioning scheme that depends on other agents' outcomes, it is equally effective to employ a modified sanctioning scheme in which the agent's sanction is conditioned only on his own outcomes, where the sanctioning scheme offers the same expected sanctioning scheme as the original contract. Second, conditional on which of his tasks fail inspection, an agent has no influence over which of his tasks pass inspection—passed inspections depend entirely on how the other agents monitor him. Specifically, fix a set of tasks that fail inspection, and suppose the agent considers completing an additional task. For monitoring realizations in which that task is monitored, he reduces the number of tasks that fail inspection. For monitoring realizations in which that task is not monitored, he does not affect how many tasks fail or pass inspection. Thus, the agent's incentives under a contract depending on both failed and passed inspections can be replicated by a contract that, conditional on failed inspections, offers the same sanction regardless

[22] Al-Najjar, Anderlini, and Felli (2006) characterize finite contracts regarding "undescribable" events, which can be fully understood only using countably infinite statements. In this interpretation, to carry out an undescribable task properly, a player must memorize and recall an infinite statement. The related literature considers contracts with bounded rationality concerns relating to complexity—such as limitations on thinking through or foreseeing contingencies (e.g., Maskin and Tirole 1999; Tirole 2009; Bolton and Faure-Grimaud 2010), communication complexity (e.g., Segal 1999), and contractual complexity (e.g., Anderlini and Felli 1998; Battigalli and Maggi 2002).

of passed inspections, where the sanctioning scheme offers the same expected sanctioning scheme as the original contract.

Part (iii): Suppose, to the contrary, that there is an optimal contract in which, when $a \leq p_i$ tasks are feasible, an agent is supposed to complete $s_i(a) < a$ tasks, but is indifferent between completing $s_i(a)$ tasks and completing $a' \leq a$ tasks, with $a' > s_i(a)$. But then there exists a superior contract, otherwise unchanged, in which he simply completes a' tasks whenever a tasks are feasible—he is no worse off himself, and his team members are strictly better off.

Part (iv): By revealed preference, $s_i(s_i(a)) = s_i(a)$, so it suffices to show that $a < a'$ implies $s_i(a) \leq s_i(a')$. Suppose to the contrary that $a < a'$ and $s_i(a) > s_i(a')$. Since upward incentive constraints are slack, the agent must strictly prefer to complete $s_i(a')$ tasks over $s_i(a)$ tasks when a' tasks are feasible. But then the agent must prefer the same when only a tasks are feasible, a contradiction to incentive compatibility of s_i.

We now allow for an imperfect monitoring technology: an uncompleted task fails inspection with probability $\gamma \in (0, 1]$. Let p_i be the number of tasks agent i is assigned, of which F_i will be monitored. If agent i fulfills a tasks, and tasks are drawn uniformly for monitoring, then the probability i has f failed inspections is given by the compound hypergeometric-binomial distribution

$$g(f, a) = \sum_{k=f}^{F_i} \frac{\binom{p_i - a}{k}\binom{a}{F_i - k}}{\binom{p_i}{F_i}} \binom{k}{f} \gamma^f (1 - \gamma)^{k-f}. \tag{A1}$$

To interpret equation (A1), observe that in order to discover f failures of agent i, the monitor(s) must have drawn $k \geq f$ tasks from the $p_i - a$ tasks agent i failed to fulfill, and $F_i - k$ tasks from the a tasks agent i fulfilled; this is described by a hypergeometric distribution. Of these k tasks, the monitor(s) must then identify exactly f failed inspections; this distribution is described by a binomial distribution. This compound hypergeometric binomial distribution is studied by Johnson and Kotz (1985) and shown by Stefanski (1992) to have a monotone likelihood ratio property: $g(f, a)/g(f, a - 1) < g(f - 1, a)/g(f - 1, a - 1)$ for all a,f. Hence, completing more tasks yields a first-order decrease in the number of failed inspections.

LEMMA 3: *The optimal way to implement full performance* (*when it gives positive welfare*) *is with linear sanctions*, $N - 2$ *agents each assigned* M *tasks*, *and two agents assigned* $M - 1$ *tasks each.*

PROOF:

By incentive-compatibility, to ensure that p_i rather than $a < p_i$ tasks are performed when p_i are feasible, we need $h_{v_i}(a) \leq h_{v_i}(p_i) + (p_i - a)\left(\frac{b}{N} - c\right)$, where

$h_{v_i}(\cdot)$ is the expected sanction conditional on the number of tasks completed. This means that $h_{v_i}(a)$ can be at best $(p_i - a)\left(\frac{b}{N} - c\right)$. We claim this can be achieved as in the statement of the lemma. Suppose each of two supervisors (agents $N-1$ and N) monitor F tasks. We divide the entire set of agents into two, each assigned to a different supervisor's responsibility for monitoring (clearly each supervisor must be assigned to the group of the other supervisor). Each supervisor randomizes uniformly over which of their agents to monitor, and then uniformly over which task of that agent to monitor. Let N_i be the number of agents in the group to which i belongs. Penalties depend on the number f of failed inspections. Let agent i's sanctioning scheme be $v_i(f) = fN_i\left(\frac{b}{N} - c\right)\frac{p_i}{\gamma F}$, so that

$$\text{(A2)} \qquad \begin{aligned} h_{v_i}(a) &= \frac{1}{N_i}\sum_{f=0}^{F} v_i(f)g(f,a) \\ &= \frac{p_i}{\gamma F}\left(\frac{b}{N} - c\right)\sum_{f=0}^{F} fg(f,a) = (p_i - a)\left(\frac{b}{N} - c\right), \end{aligned}$$

because the expectation of the compound hypergeometric-binomial is $(p_i - a)\gamma F/p_i$. Conditional expected sanctions are independent of F, for $F \geq 1$. This contract gives expected social utility

$$\begin{aligned} \text{(A3)} \quad & \sum_{i=1}^{N}\sum_{a=0}^{M}\binom{p_i}{a}\lambda^a(1-\lambda)^{p_i-a}\left((b-c)a + (p_i - a)\left(\frac{b}{N} - c\right)\right) \\ &= \left(\frac{b}{N} - c\right)\sum_{i=1}^{N} p_i\sum_{a=0}^{M}\binom{p_i}{a}\lambda^a(1-\lambda)^{p_i-a} + \frac{N-1}{N}b\sum_{a=0}^{p_i}\binom{p_i}{a}\lambda^a(1-\lambda)^{p_i-a}a \\ &= \left(\frac{N-1}{N}b\lambda + \frac{b}{N} - c\right)\sum_{i=1}^{N} p_i. \end{aligned}$$

This is positive if $\lambda > \left(c - \frac{b}{N}\right)/\frac{N-1}{N}b$ and largest when the maximal number of tasks are assigned, using $F = 1$ for each of the two supervisors: $p_i = M$ for $i = 1, 2, \ldots, N-2$ and $p_{N-1} = p_N = M - 1$.

LEMMA 4: *Suppose agent i is assigned M tasks, of which F_i are monitored. If $v_i(f) = 0$ for $f < F_i$ and $v_i(F_i) < 0$, then a work target strategy is induced. If $v_i(F_i)$ is chosen optimally conditional on p_i^* being the induced work target, then agent i's contribution to social welfare given by*

$$\text{(A4)} \quad \sum_{a=0}^{M}\binom{M}{a}\lambda^a(1-\lambda)^{M-a}\left((b-c)\min\{a, p_i^*\} + \frac{(c - \frac{b}{N})g(F_i, \min\{a, p_i^*\})}{g(F_i, p_i^*) - g(F_i, p_i^* - 1)}\right).$$

The value of equation (A4) *is strictly increasing and concave in λ. Moreover, if $p_i^* < \tilde{p}_i^* \leq M - F + 1$, the value of equation* (A4) *for $\tilde{p}_i^*$ strictly single crosses the value of equation* (A4) *for p_i^* from below, as a function of λ.*

PROOF:

That a cutoff strategy is induced follows from the first paragraph in the proof of Theorem 4 (in the online Appendix). Given p_i^*, equation (A4) follows from choosing $v_i(F_i)$ to make the incentive constraint for doing p_i^* versus $p_i^* - 1$ tasks bind. Let

$$\beta(a) \equiv (b-c)\min\{a, p_i^*\} + \frac{\left(c - \frac{b}{N}\right) g(F, \min\{a, p_i^*\})}{g(F, p^*) - g(F, p^* - 1)}. \tag{A5}$$

Equation (A4) is the expectation of $\beta(a)$ with respect to the binomial distribution. The first term of $\beta(a)$ is concave in a. Moreover,

$$\begin{aligned}
&\binom{M - \min\{a+1, p_i^*\}}{F} - 2\binom{M - \min\{a, p_i^*\}}{F} + \binom{M - \min\{a-1, p_i^*\}}{F} \\
&\quad = \begin{cases} \binom{M-a}{F}\left(\dfrac{F}{M-(a+1)-F} - \dfrac{F}{M-a}\right) & \text{if } a \le p_i^* - 1, \\ \binom{M-(p_i^*-1)}{F} - \binom{M-p_i^*}{F} & \text{if } a = p_i^*, \\ 0 & \text{if } a \ge p_i^* + 1. \end{cases}
\end{aligned}$$

which is positive because $F \ge 1$, and $M - p_i^* + 1 \ge F$. Hence

$$g(F, \min\{a, p_i^*\}) = \lambda^F \frac{\binom{M - \min\{a, p_i^*\}}{F}}{\binom{M}{F}} \tag{A7}$$

is convex. The second term in $\beta(a)$ is a negative constant times $g(F, \min\{a, p_i^*\})$, so $\beta(a)$ is concave. Finally, the binomial distribution satisfies double-crossing, since

$$\begin{aligned}
&\frac{\partial^2}{\partial \lambda^2}\left(\binom{M}{a}\lambda^a(1-\lambda)^{M-a}\right) \\
&\quad = \binom{M}{a}(1-\lambda)^{M-2-a}\lambda^{a-2}(a^2 - (1 + 2(M-1)\lambda)a + M(M-1)\lambda^2)
\end{aligned} \tag{A8}$$

is negative if and only if $a^2 - (1 + 2(M-1)\lambda)a + M(M-1)\lambda^2 < 0$. Hence, by Lemma 5 (in the online Appendix), equation (A4) is concave in λ. To see that is increasing in λ, observe that the benefit of each task is linear in a, increasing in p_i^* and independent of λ (a parameter of first-order stochastic dominance for the binomial distribution). The expected sanction for completing $\min\{a, p_i^*\}$ tasks is $\left(\left(c - \frac{b}{N}\right) g(F, \min\{a, p_i^*\})\right) / (g(F, p_i^*) - g(F, p_i^* - 1))$. As λ cancels out of the above, we need only check this expression has increasing differences in a and p_i^* (by

Corollary 10 of Van Zandt and Vives 2007). Denote a p_i^*-work-target strategy by $s_{p_i^*}$. Since $c > \frac{b}{N}$, the sign of the second difference depends on

$$\text{(A9)}\quad \frac{g(F, s_{p_i^*+1}(a+1)) - g(F, s_{p_i^*+1}(a))}{g(F, p_i^* + 1) - g(F, p_i^*)} - \frac{g(F, s_{p_i^*}(a+1)) - g(F, s_{p_i^*}(a))}{g(F, p_i^*) - g(F, p_i^* - 1)}$$

$$= \begin{cases} 0 & \text{if } a \geq p_i^* + 1, \\ 1 & \text{if } a = p_i^*, \\ \dfrac{g(F, a+1) - g(F, a)}{g(F, p_i^* + 1) - g(F, p_i^*)} - \dfrac{g(F, a+1) - g(F, a)}{g(F, p_i^*) - g(F, p_i^* - 1)} & \text{if } a \leq p_i^* - 1. \end{cases}$$

Concentrating on the third case, since $g(F, a)$ is decreasing in a, it suffices to show that

$$\text{(A10)}\quad \binom{M - p_i^*}{F} - \binom{M - p_i^* + 1}{F} > \binom{M - p_i^* + 1}{F} - \binom{M - p_i^* + 2}{F}.$$

But this is exactly analogous to the earlier calculation.

PROOF OF THEOREM 3 (Continued):

It remains to show the second statement. When $N = 2$, applies, so suppose $N > 2$. By the De Moivre–Laplace theorem (Johnson, Kemp, and Kotz 2005 equation 3.20), the normal distribution with mean $p\lambda$ and variance $p(1 - \lambda)\lambda$ approximates the binomial distribution with p tasks, each with λ probability of being feasible, and the approximation error in the CDF at any point is $O(p\lambda(1 - \lambda))^{1/2}$. Using this approximation, we define the *continuous problem* of choosing to assign $\tilde{p}_i \in \mathbb{R}$ tasks and work target $\tilde{p}_i^* \in \mathbb{R}$ to solve

$$\text{(A11)}\quad \max_{\{\tilde{p}_i, \tilde{p}_i^* \in \mathbb{R}\}_{i=1}^N} \sum_{i=1}^n E((b - c)\min\{a, \tilde{p}_i^*\} + \left(\frac{b}{N} - c\right)\max\{\tilde{p}_i^* - a, 0\})$$

$$\text{s.t. } \tilde{p}_i^* \leq \tilde{p}_i \leq M \text{ for all } i \text{ and } \sum_i \tilde{p}_i \leq M(N - 1),$$

where expectation of $\tilde{p}_i^*$ is taken with respect to the normal distribution $\mathcal{N}(\tilde{p}_i\,\lambda, \tilde{p}_i(1 - \lambda)\lambda)$. Write the objective function as $\sum_i E_i$, where $E_i \equiv E\Big((b - c)\min\{a, \tilde{p}_i^*\} + \left(\frac{b}{N} - c\right)\max\{\tilde{p}_i^* - a, 0\}\Big)$. We begin with the inner part of this problem—optimizing $\tilde{p}_i^*$ given $\tilde{p}_i$. The first-order condition is

$$\text{(A12)}\quad \frac{\partial E_i}{\partial \tilde{p}_i^*} = \frac{1}{2N}\left(b + bN - 2cN + b(N - 1)\,\mathrm{erf}\left(\frac{\lambda\tilde{p}_i - \tilde{p}_i^*}{\sqrt{2\tilde{p}(1 - \lambda)\lambda}}\right)\right) = 0,$$

where $\text{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$. The first order condition is solved at

$$\text{(A13)} \qquad \tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i) \equiv \tilde{p}_i\,\lambda - \sqrt{2\tilde{p}_i(1-\lambda)}\lambda\ \text{erf}^{-1}\left(\frac{b + bN - 2cN}{b - bN}\right)$$

and the welfare arising from each agent i is

$$\text{(A14)} \qquad \tilde{p}_i\lambda(c - b) - be^{-\text{erf}^{-1}\left(\frac{b+bN-2cN}{b-bN}\right)^2} \frac{N-1}{N}\sqrt{\frac{\tilde{p}_i\lambda(1-\lambda)}{2\pi}}.$$

The strict second-order condition is satisfied globally:

$$\text{(A15)} \qquad \frac{\partial^2 E_i}{\partial(\tilde{p}_i^{\,*})^2} = -be^{-\frac{(\lambda\tilde{p}_i - \tilde{p}_i^{\,*})^2}{2\tilde{p}(1-\lambda)\lambda}} \frac{N-1}{N}\sqrt{\frac{1}{2\pi\tilde{p}(1-\lambda)\lambda}} < 0.$$

We now move to the outer part of the continuous problem, choosing p_i. By the envelope theorem, $\frac{dE_i}{d\tilde{p}_i}\Big|_{\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)} = \frac{\partial E_i}{\partial \tilde{p}_i}\Big|_{\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)}$ and $\frac{d^2E_i}{d\tilde{p}_i^2}\Big|_{\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)} = \left(\frac{\partial^2 E_i}{\partial \tilde{p}_i^2} + \frac{\partial^2 E_i}{\partial p_i \partial \tilde{p}_i^{\,*}}\frac{d\tilde{P}^*}{d\tilde{p}_i)}\right)\Big|_{\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)}$. Solving the closed form of $\frac{d^2E_i}{dp_i^2}$ at $\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)$ shows the objective is strictly convex in each $\tilde{p}_i$:

$$\text{(A16)} \qquad \frac{d^2E_i}{dp_i^2}\Big|_{\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)} = \frac{1}{4\tilde{p}_i^{\,2}}\, be^{-\text{erf}^{-1}\left(\frac{b+bN-2cN}{N-bN}\right)^2} \frac{N-1}{N}\sqrt{\frac{\tilde{p}_i\lambda(1-\lambda)}{2\pi}} > 0.$$

Since, in addition, $\frac{dE_i}{d\tilde{p}_i} > 0$ at $\tilde{p}_i^{\,*} = \tilde{P}^*(\tilde{p}_i)$, and M units of monitoring requires $\sum_i \tilde{p}_i \le (N-1)M$, it follows that the optimal assignment scheme in the continuous problem is for $N-1$ agents each to be assigned $\tilde{p}_i = M$ tasks and complete $\tilde{p}^* \equiv \tilde{P}^*(M)$, while the Nth agent only monitors. We construct a contract for the true (discrete) model, using the same assignment scheme: $N-1$ agents have M tasks, while one agent supervises. The work target must be an integer, so we round $\tilde{p}^*$ up to $\lceil \tilde{p}^* \rceil$. Let $\hat{V}$ be the welfare attained by this discrete contract, and let $\tilde{V}$ be the value of the continuous problem. The difference $\hat{V} - \tilde{V}$ arises from four issues:

- The "tail benefit": The discrete contract applies to a distribution with a lower bound of zero feasible tasks, and so does not involve the harsh sanctions that arise in the long lower tail of the continuous problem.
- The "integer benefit": The maximum number of tasks accomplished is greater under the discrete contract than in the solution to the continuous problem, leading to higher social payoffs for realizations with many feasible tasks.
- The "integer deficit": Because only whole tasks can be performed under the discrete contract, when fewer than $\lceil \tilde{p}^* \rceil$ tasks are performed the actual sanctions may be harsher than in the solution to the continuous problem.

- Approximation error: The CDF of the normal distribution at $a + \frac{1}{2}$ is only an approximation of the binomial CDF at a.

Let $\delta = \frac{N-1}{N} b$ and $\rho = (b - c)$. Let Φ and ϕ be the CDF and PDF of the normal distribution, and Ψ and ψ be the CDF and PDF of the binomial. The tail benefit (which is not affected by approximation error) is

$$X = -\int_{-\infty}^{-\frac{1}{2}} ((\delta - \rho)(\lceil \tilde{p}^* \rceil - \tilde{p}^*) + \delta a)\phi(a)\, da. \tag{A17}$$

The integer benefit, accounting for approximation error, is at least

$$Y = \rho\Big((1 - \Psi(\lceil \tilde{p}^* \rceil - 1))\lceil \tilde{p}^* \rceil - \Big(1 - \Phi\Big(\lceil \tilde{p}^* \rceil - \frac{1}{2}\Big)\Big)\tilde{p}^*\Big). \tag{A18}$$

The integer deficit, accounting for approximation error, is

$$Z = \sum_{a=0}^{\lceil \tilde{p}^* \rceil - 1} \Big(\int_{a-\frac{1}{2}}^{a+\frac{1}{2}} ((\delta - \rho)(\lceil \tilde{p}^* \rceil - \tilde{p}^*) + \delta \tilde{a})\phi(\tilde{a})\, d\tilde{a} - \delta a \psi(a) \Big). \tag{A19}$$

Combining terms and collecting $\lceil \tilde{p}^* \rceil - \tilde{p}^*$ gives the welfare deficit from each of the $N - 1$ task-performing agents under the discrete contract, compared to the value of the continuous problem:

$$-\frac{1}{N-1}(\hat{V} - \tilde{V}) = Z - X - Y \tag{A20}$$

$$= -(\lceil \tilde{p}^* \rceil - \tilde{p}^*)\Big(\rho - \delta\Phi\Big(\lceil \tilde{p}^* \rceil - \frac{1}{2}\Big)\Big) - \rho\Big(\Phi\Big(\lceil \tilde{p}^* \rceil - \frac{1}{2}\Big) - \Psi(\lceil \tilde{p}^* \rceil - 1)\Big)\lceil \tilde{p}^* \rceil$$

$$- \sum_{a=0}^{\lceil \tilde{p}^* \rceil - 1} \delta a \psi(a) + \int_{-\infty}^{\lceil \tilde{p}^* \rceil - \frac{1}{2}} \delta a \phi(a)\, da.$$

The right-hand side of the first line is bounded by $[\rho - \delta, \rho]$ regardless of $\tilde{p}^*$, while the terms on the second line are on the order of $\tilde{p}^*$ times the approximation error between Φ and Ψ. By the De Moivre–Laplace theorem, the approximation error is on the order of $M^{-1/2}$. Since by equation (A13) and equation (A14) both $\tilde{p}^*$ and the value of the continuous problem are on the order of M, the ratio of the welfare under the discrete contract and the value of the continuous problem converges to 1 as $M \to \infty$. Consider the true optimal contract in the discrete problem. Let p_i be the number of tasks assigned to agent i and p_i^* be i's work target. This contract's value can be approximated by evaluating the objective of the continuous problem at $p_i + \frac{1}{2}$ and $p_i^* + \frac{1}{2}$. By a similar argument, the deficit per agent of this approximation compared to the true value of the optimal discrete contract is no more than the following term, which is on the order of $M^{1/2}$:

$$\frac{1}{\mathrm{N}} \sum_i \Bigg[\gamma\Bigg(\Phi\Big(\lceil \tilde{p}^* \rceil - \frac{1}{2} \Big) - \Psi(\lceil \tilde{p}^* \rceil - 1) \Bigg) \lceil \tilde{p}^* \rceil \tag{A21}$$

$$+ \sum_{a=0}^{\lceil \tilde{p}^* \rceil - 1} \delta a \psi(a) - \int_{-\infty}^{\lceil \tilde{p}^* \rceil - \frac{1}{2}} \delta a \phi(a)\, da \Bigg].$$

Thus, the ratio of $\hat{V}$ and the value of true optimal contract converges to 1 as $M \to \infty$.

PROOF OF THEOREM 4:

For part (i), observe that at $\lambda = 1$, in every optimal contract each of $N - 2$ agents must be assigned M tasks, and two minimal supervisors are each assigned $M - 1$ tasks, with all agents fulfilling all of them. The contract must impose harsh enough sanctions to make it incentive compatible for them to do so, but the sanctions may be arbitrarily severe since they are not realized on the equilibrium path. The value of any such contract is $(NM - 2)(b - c)$. Fix M, b, and c. The value of a contract is continuous in λ, p, s, and v. Both p and s are defined on compact spaces, and v can without loss of generality take values from the extended nonpositive real numbers $[-\infty, 0]$. Since the incentive constraints are weak inequalities that are continuous in λ, b, c, P, v, and s, the constraint set is compact-valued for each λ, b, and c. The constraint set is nonempty, as it always contains the contract in which no tasks are assigned and no sanctions are imposed. By Berge's Theorem of the Maximum (Aliprantis and Border 2006, Theorem 17.31), the value of an optimal contract is continuous in λ and the correspondence mapping λ to the set of optimal contracts (v, s, and ρ) is upper hemicontinuous. The value of the contract must converge to $(NM - 2)(b - c)$ as $\lambda \to 1$, and so must have the same number of tasks assigned per agent as above for λ sufficiently high. To minimize sanction costs, all downward constraints for completing that number of tasks should bind, which is achieved by a linear contract with uniform randomization over monitored tasks. Given a linear contract, $s(a) = a$ for all a is optimal. Now apply Lemma 3.

For part (ii), we show underperformance with M monitoring units strictly dominates full performance at $\lambda^* = \frac{cN - b}{(N-1)b}$, and, by single crossing, for all $\lambda \leq \lambda^*$. Since $c < b < cN$, $\lambda^* \in (0, 1)$. The value of full performance, $(MN - 2)\left(\frac{N-1}{N} b\lambda + \frac{b}{N} - c\right)$, is zero at λ^*. The value of a single-supervisor arrangement with a work target of $M - 1$ for workers is

$$(N-1)\Bigg((M-1)\left(\frac{b}{N} - c\right) + \frac{N-1}{N} b \sum_{a=0}^{M} \binom{M}{a} \lambda^a (1-\lambda)^{m-a} \min\{a, M-1\}\Bigg), \tag{A22}$$

which simplifies to

$$(N-1)\left((M-1)\left(\frac{b}{N} - c\right) + \frac{N-1}{N} b(M\lambda - \lambda^M)\right). \tag{A23}$$

This equals zero at the solution $\hat{\lambda}$ to $\frac{M\lambda - \lambda^M}{M-1} = \lambda^*$. By Descartes' Rule of Signs, the only real roots of $\frac{M\lambda - \lambda^M}{M-1} = \lambda$ are $\lambda = 0$ and $\lambda = 1$, and

$\frac{d}{d\lambda}\left(\frac{M\lambda-\lambda^M}{M-1}\right)\Big|_{\lambda=0} = \frac{M}{M-1} > 1$. These facts imply the solution $\hat{\lambda}$ satisfies $\hat{\lambda} < \lambda^*$. Since the contract's value strictly increases in λ, underperformance with a single supervisor dominates full performance on a neighborhood of λ^*.

Equation (4) implies that the value of work target contracts with a single supervisor, as a function of λ, is a single crossing family in p^*. This family includes a full performance contract with a single supervisor, although this full performance contract is dominated by the optimal full performance contract with two minimal supervisors. By the single crossing property, underperformance with a single supervisor dominates the suboptimal full performance contract with a single supervisor for all $\lambda \in [0, \lambda^*]$. Moreover, the optimal full performance contract single crosses the suboptimal full performance contract from below (observe that both are linear, while the optimal full performance contract results in harsher sanctions at $\lambda = 0$). Since the underperformance contract with a single supervisor dominates the optimal full performance contract at λ^*, it also does so for all $\lambda \in [0, \lambda^*]$.

Finally, for part (iii), we show that when $\frac{c}{b} > \frac{2}{e+2} + \frac{1}{N}\frac{e}{e+2}$, underperformance strictly better than full performance for $\lambda = \frac{M-2}{M-1}$, and thus (by continuity) for an open neighborhood. Consider a maximally forgiving contract against the $N-2$ workers to enforce cutoff p^* as in Lemma 4. Combining Lemma 3 and Lemma 4 for $F = 2$, the contract's social value is

$$\text{(A24)}\quad 2(M-1)\left(\frac{N-1}{N}b\lambda + \frac{b}{N} - c\right)$$
$$+ (N-2)\sum_{a=0}^{M}\binom{M}{a}\lambda^a(1-\lambda)^{M-a}\left(\min\{a,p^*\}(b-c) + \frac{\frac{b}{N}-c}{M-p^*}\binom{M-\min\{a,p^*\}}{2}\right).$$

By Lemma 4 this is a single-crossing family and the optimal p^* increases with λ. Note that $\sum_{a=0}^{M}\binom{M}{a}\lambda^a(1-\lambda)^{M-a}a^2 = M\lambda(1-\lambda) + (M\lambda)^2$. When $p^* = M-1$, reduces to

$$\text{(A25)}\quad 2(M-1)\left(\frac{N-1}{N}b\lambda + \frac{b}{N} - c\right)$$
$$+ (N-2)\left((M\lambda-\lambda^M)(b-c) + \left(\frac{b}{N}-c\right)\frac{M(M-1)(1-\lambda)^2}{2}\right).$$

Simplifying terms and factoring, equation (A25) dominates full performance when

$$\text{(A26)}\quad 2N(M\lambda-\lambda^M)(b-c) + (b-cN)(M(M-1)(1-\lambda)^2)$$
$$> 2M((N-1)b\lambda + b - cN)$$
$$\Leftrightarrow\ 2(b-c)\lambda^M N < (1-\lambda)(M-\lambda(M-1)-3)M(b-cN).$$

The last LHS in equation (A26) is positive, so the RHS must be positive too, for equation (A26) to hold. So $b < cN$ means $M - \lambda(M-1) - 3 < 0$. For $\lambda^* = \frac{M-2}{M-1}$, equation (A26) is

$$2(b-c)\left(\frac{M-2}{M-1}\right)^M N < \frac{M}{M-1}(cN-b). \tag{A27}$$

The RHS of equation (A27) is bounded below by $cN - b$. Consider $z_M = \left(\frac{M-2}{M-1}\right)^M$. Taking logarithms, $\ln z_M = M \ln\frac{M-2}{M-1}$. Using l'Hôpital's rule, $\lim_{M\to\infty} \ln z_M = \lim_{M\to\infty} \frac{-M^2}{(M-2)(M-1)} = -1$, hence, $z_M \to \frac{1}{e}$ from below. The LHS of equation (A27) is then bounded above by $\frac{2(b-c)N}{e}$, and a sufficient condition for equation (A27) is $\frac{2(b-c)N}{e} < cN - b$, which rearranges to $\frac{c}{b} > \frac{2}{e+2} + \frac{1}{N}\frac{e}{e+2}$.

REFERENCES

Adams, Scott. 1996. *The Dilbert Principle: A Cubicle's-Eye View of Bosses, Meetings, Management Fads & Other Workplace Afflictions*. New York: HarperCollins Publishers, Inc.

Alchian, Armen A., and Harold Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 62 (5): 777–95.

Aliprantis, Charalambos D., and Kim C. Border. 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. New York: Springer Berlin Heidelberg.

Al-Najjar, Nabil I., Luca Anderlini, and Leonardo Felli. 2006. "Undescribable Events." *Review of Economic Studies* 73 (4): 849–68.

Anderlini, Luca, and Leonardo Felli. 1998. "Describability and Agency Problems." *European Economic Review* 42 (1): 35–59.

Aoki, Masahiko. 1988. *Information, Incentives, and Bargaining in the Japanese Economy*. New York: Cambridge University Press.

Aoki, Masahiko. 1994. "The continent governance of teams: Analysis of institutional complementarity." *International Economic Review* 35 (3): 657–76.

Aron, Debra J., and Pau Olivella. 1994. "Bonus and Penalty Schemes as Equilibrium Incentive Devices, with Application to Manufacturing Systems." *Journal of Law, Economics, and Organization* 10 (1): 1–34.

Baker, George P., Michael C. Jensen, and Kevin J. Murphy. 1988. "Compensation and Incentives: Practice vs. Theory." *Journal of Finance* 43 (3): 593–616.

Barron, John M., and Kathy Paulson Gjerde. 1997. "Peer Pressure in an Agency Relationship." *Journal of Labor Economics* 15 (2): 234–54.

Battaglini, Marco. 2006. "Joint Production in Teams." *Journal of Economic Theory* 130 (1): 138–67.

Battigalli, Pierpaolo, and Giovanni Maggi. 2002. "Rigidity, Discretion, and the Costs of Writing Contracts." *American Economic Review* 92 (4): 798–817.

Bénabou, Roland. 2013. "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies* 80 (2): 429–62.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (3): 871–915.

Bodoh-Creed, Aaron L. 2013. "Mood, Memory, and the Evaluation of Asset Prices." http://faculty.haas.berkeley.edu/acreed/Associative%20Memory.pdf.

Bolton, Patrick, and Antoine Faure-Grimaud. 2010. "Satisficing Contracts." *Review of Economic Studies* 77 (3): 937–71.

Boning, Brent, Casey Ichniowski, and Kathryn Shaw. 2007. "Opportunity Counts: Teams and the Effectiveness of Production Incentives." *Journal of Labor Economics* 25 (4): 613–50.

Border, Kim C., and Joel Sobel. 1987. "Samurai Accountant: A Theory of Auditing and Plunder." *Review of Economic Studies* 54 (4): 525–40.

Carpenter, Jeffrey, Samuel Bowles, Herbert Gintis, and Sung-Ha Hwang. 2009. "Strong Reciprocity and Team Production: Theory and Evidence." *Journal of Economic Behavior & Organization* 71 (2): 221–32.

Chakraborty, Archishman, and Rick Harbaugh. 2007. "Comparative Cheap Talk." *Journal of Economic Theory* 132 (1): 70–94.

Che, Yeon-Koo, and Seung-Weon Yoo. 2001. "Optimal Incentives for Teams." *American Economic Review* 91 (3): 525–41.

Cole, Harold L., and Narayana R. Kocherlakota. 2005. "Finite Memory and Imperfect Monitoring." *Games and Economic Behavior* 53 (1): 59–72.

Compte, Olivier, and Andrew Postlewaite. 2008. "Repeated Relationships with Limits on Information Processing." University of Pennsylvania Working Paper 08-026.

Coviello, Decio, Andrea Ichino, and Nicola Persico. 2014. "Time Allocation and Task Juggling." *American Economic Review* 104 (2): 609–23.

Cowan, Nelson. 2000. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24 (1): 87–114.

d'Aspremont, Claude, and Louis-André Gérard-Varet. 1998. "Linear Inequality Methods to Enforce Partnerships under Uncertainty: An Overview." *Games and Economic Behavior* 25 (2): 311–36.

Dow, James. 1991. "Search Decisions with Limited Memory." *Review of Economic Studies* 58 (1): 1–14.

Fortney, Susan Saab. 1995. "Am I My Partner's Keeper? Peer Review in Law Firms." *University of Colorado Law Review* 66 (2): 329–73.

Frankel, Alexander. 2014. "Aligned delegation." *American Economic Review* 104 (1): 66–83.

Frederickson, James R., and William Waller. 2005. "Carrot or Stick? Contract Frame and Use of Decision-Influencing Information in a Principal-Agent Setting." *Journal of Accounting Research* 43 (5): 709–33.

Hirshleifer, David, and Ivo Welch. 2004. "An Economic Approach to the Psychology of Change: Amnesia, Inertia, and Impulsiveness." *Journal of Economics and Management Strategy* 11 (3): 379–421.

Holmstrom, Bengt. 1982. "Moral hazard in teams." *Bell Journal of Economics* 13 (2): 324–40.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics & Organization* 7 (Special Issue): 24–52.

Jackson, Matthew O., and Hugo F. Sonnenschein. 2007. "Overcoming Incentive Constraints by Linking Decisions." *Econometrica* 75 (1): 241–57.

Johnson, Norman L., Adrienne W. Kemp, and Samuel Kotz. 2005. *Univariate Discrete Distributions*. 3rd ed. Hoboken, New Jersey: John Wiley & Sons.

Johnson, Norman L., and Samuel Kotz. 1985. "Some Distributions Arising as a Consequence of Errors in Inspection." *Naval Research Logistics Quarterly* 32 (1): 35–43.

Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure and Partnerships." *Journal of Political Economy* 100 (4): 801–17.

Knez, Marc, and Duncan Simester. 2001. "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines." *Journal of Labor Economics* 19 (4): 743–72.

Kocer, Yilmaz. 2012. "Endogenous Learning with Bounded Memory." http://www.usc.edu/schools/business/FBE/seminars/papers/AE_10-12-12_KOCER.pdf.

Kvaløy, Ola, and Trond E. Olsen. 2006. "Team Incentives in Relational Employment Contracts." *Journal of Labor Economics* 24 (1): 139–69.

Laux, Christian. 2001. "Limited Liability and Incentive Contracting with Multiple Projects." *RAND Journal of Economics* 32 (3): 514–26.

Lazear, Edward P., and Kathryn L. Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives* 21 (4): 91–114.

Legros, Patrick, and Hitoshi Matsushima. 1991. "Efficiency in Partnerships." *Journal of Economic Theory* 55 (2): 296–322.

Legros, Patrick, and Steven A. Matthews. 1993. "Efficient and Nearly-Efficient Partnerships." *Review of Economic Studies* 60 (3): 599–611.

Li, Shuhe, and Weiying Zhang. 2001. "Optimal Assignment of Principalship in Teams." *Journal of Economic Behavior & Organization* 44 (1): 105–27.

Lorentz, G. G. 1953. "An Inequality for Rearrangements." *American Mathematical Monthly* 60 (3): 176–79.

Maskin, Eric, and Jean Tirole. 1999. "Unforeseen Contingencies and Incomplete Contracts." *Review of Economic Studies* 66 (1): 83–114.

Matsushima, Hitoshi, Koichi Miyazaki, and Nobuyuki Yagi. 2010. "Role of Linking Mechanisms in Multitask Agency with Hidden Information." *Journal of Economic Theory* 145 (6): 2241–59.
McAfee, R. Preston, and John McMillan. 1991. "Optimal Contracts for Teams." *International Economic Review* 32 (3): 561–77.
Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information." *Psychological Review* 63 (2): 81–97.
Miller, Nolan H. 1997. "Efficiency in Partnerships with Joint Monitoring." *Journal of Economic Theory* 77 (2): 285–99.
Mirrlees, James A. 1976. "The Optimal Structure of Incentives and Authority within an Organization." *Bell Journal of Economics* 7 (1): 105–31.
Mookherhjee, Dilip, and Ivan Png. 1989. "Optimal Auditing, Insurance, and Redistribution." *Quarterly Journal of Economics* 104 (2): 399–415.
Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 117 (3): 735–74.
Oyer, Paul, and Scott Schaefer. 2005. "Why Do Some Firms Give Stock Options to All Employees?: An Empirical Examination of Alternative Theories." *Journal of Financial Economics* 76 (1): 99–133.
Palfrey, Thomas R., and Howard Rosenthal. 1984. "Participation and the Provision of Discrete Public Goods: A Strategic Analysis." *Journal of Public Economics* 24 (2): 171–93.
Piccione, Michele, and Ariel Rubinstein. 1993. "Finite Automata Play a Repeated Extensive Game." *Journal of Economic Theory* 61 (1): 160–68.
Piccione, Michele, and Ariel Rubinstein. 1997. "On the Interpretation of Decision Problems with Imperfect Recall." *Games and Economic Behavior* 20 (1): 3–24.
Rahman, David. 2012. "But Who Will Monitor the Monitor?" *American Economic Review* 102 (6): 2267–97.
Rahman, David, and Ichiro Obara. 2010. "Mediated Partnerships." *Econometrica* 78 (1): 285–308.
Richmond, Douglas R. 2007–08. "Law Firm Partners as their Brothers' Keepers." *Kentucky Law Journal* 96: 231–73.
Romero, Julian. 2011. "Finite Automata in Undiscounted Repeated Games with Private Monitoring." https://www.krannert.purdue.edu/programs/phd/working-papers-series/2011/1260.pdf.
Segal, Ilya. 1999. "Complexity and Renegotiation: A Foundation for Incomplete Contracts." *Review of Economic Studies* 66 (1): 57–82.
Snyder, Christopher M. 1999. "Bounding the Benefits of Stochastic Auditing: The Case of Risk-Neutral Agents." *Economic Theory* 14 (1): 247–53.
Stefansk, Leonard. 1992. "Monotone Likelihood Ratio of a 'Faulty-Inspection' Distribution." *American Statistician* 46 (2): 110–14.
Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *American Economic Review* 99 (1): 265–94.
Townsend, Robert M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (2): 265–93.
Van Zandt, Timothy, and Xavier Vives. 2007. "Monotone Equilibria in Bayesian Games of Strategic Complementarities." *Journal of Economic Theory* 134 (1): 339–60.
Williamson, Stephen D. 1987. "Costly Monitoring, Loan Contracts, and Equilibrium Credit Rationing." *Quarterly Journal of Economics* 102 (1): 135–45.
Wilson, Andrea. 2014. "Bounded Memory and Biases in Information Processiong." https://sites.google.com/site/andreamaviwilson/research/memorybiases_round2_WITH-EXAMPLE.pdf.