Simple Incentives and Diverse Beliefs

Maxwell Rosenthal*

September 25, 2025

Abstract

This paper studies a moral hazard problem in which the principal does not know the agent's beliefs about the output generating process. The agent is risk neutral, transfers are subject to limited liability, and the principal evaluates contracts according to their worst-case payoff against a rich set of plausible agent beliefs. With common knowledge of the relationship between effort and expected output, optimal contracts are of the form $w(y) = \max(\alpha y + \beta, 0)$. With or without common knowledge of that relationship, there are broad conditions under which optimal contracts are of the form $w(y) = \alpha y + \beta$ and the principal can not improve her payoff guarantee by randomizing over menus of contracts.

JEL classification: D81, D82, D86

Keywords: uncertain beliefs, robustness, linear bonus contracts

1 Introduction

In the classical formulation of the moral hazard problem, the agent's effort determines the distribution of output to be received by the principal. Because effort is costly for

*Georgia Institute of Technology. Elements of the premise of this paper emerged from conversations with Asaf Plan while the author was a graduate student at the University of Arizona. Thanks to the handling editor Bruno Strulovici and three referees for providing generously detailed feedback, as well as to Sylvain Chassang and Bart Lipman for comments on earlier versions of the paper. This work has benefited from conversations with Gabriel Carroll, Ashwin Kambhampati, and seminar participants at the Arizona State University Theory Conference and the University of Georgia.

the former and unobserved by the latter, the principal pays the agent on the basis of the output he produces. The joint distribution of effort and output is common knowledge, and these shared beliefs are inputs into both the agent's problem of choosing how hard to work and also the principal's problem of identifying the contract that most cheaply incentivizes the agent to exert her preferred level of effort.

While the common prior assumption lends a simplicity to the above framework that has proven valuable for developing principles, it is unclear whether or not its demanding informational criteria are likely to be met in practice. First, because the agent's actions are hidden from the principal, it is not obvious what might constitute an objective source for shared beliefs about the output generating process. Second, while classical decision theory (Savage (1954)) provides foundations for expected utility maximization with respect to subjective beliefs, those beliefs are personal to the decision maker (Morris (1995); Kreps (2012)). And if beliefs are indeed personal, it seems they might also be private information.

This paper studies a general moral hazard problem in which the principal does not know the agent's beliefs A about how output y is distributed. Instead, she knows only the identity of some rich set of plausible beliefs to which they belong. The agent is risk neutral, protected by limited liability, and suffers additive disutility from effort e. His action set is one-dimensional, effort costs are common knowledge, and the principal assesses her payoffs with respect to her own subjective beliefs about output. Because she does not know the agent's beliefs, the principal does not necessarily have a complete understanding of how he will respond to any particular incentive scheme. In the face of that uncertainty, she seeks a contract w that maximizes her guaranteed payoff against the set of plausible beliefs.

Our results are as follows. In Section 3, we initiate our study of the principal's problem by supposing the relationship between effort and expected output is common knowledge. First, we show optimal contracts for incentivizing uniform effort[1] across the set of plausible beliefs are linear contracts of the form $w(y) = \alpha y + \beta$ with minimum payment 0 (Theorem 1) . Second, we show optimal contracts in general are linear bonus contracts of the form $w(y) = \max(\alpha y + \beta, 0)$ with minimum payment 0

[1]Contract w *uniformly incentivizes* effort e if the agent's expected payoff (net of additive effort disutility) under contract w, belief A, and effort e matches or exceeds his expected payoff under contract w, belief A, and effort e' for every alternative effort e' and for every plausible belief A.

(Theorem 2). Third, and finally, we show that if the range of outputs is sufficiently broad then the principal's problem has a saddle point and optimal contracts are linear (Theorem 3). In all such cases, the principal can not improve her payoff guarantee by screening the agent with a menu or by randomizing over contracts.

In Section 4, we relax the assumption that the relationship between effort and expected output is common knowledge and develop broad generalizations of our main result. First, we show that if the principal knows the agent's beliefs about expected output but those beliefs differ from her own then the principal's problem has a saddle point and optimal contracts are linear (Theorem 4). This result requires no new assumptions and is a proper generalization of our third theorem. Second, we show that if the principal is instead uncertain about the agent's beliefs about expected output then the same conclusions obtain as long as her utility function is linear in output and beliefs are monotone and supermodular (Theorem 5).

To develop intuition, we illustrate Theorems 1–3 with three examples. First, regarding Theorem 1, consider the setting with common knowledge of the relationship between effort and expected output and note that the principal is certain about the agent's payoffs if and only if she pays him with a contract under which wages vary affinely with output. Given that linear contracts are optimal in other models of robust contracting in part because they equate payoffs across classes of environments (Carroll (2015)), it seems natural that such contracts might be robustly optimal in our model as well. While our first theorem shows only that linear contracts are optimal within the class of incentive schemes that assign the same level of effort to every type of agent, there are cases of our model in which that is a feature of generally optimal contracts.

Example 1. The principal's goal is to maximize guaranteed expected net output $y - w(y)$. Output lies in the interval $[0, 160]$ and the agent chooses between low effort at cost 0, medium effort at cost 8, and high effort at cost 20. The principal believes output is 40 with probability 1 under low effort, 80 with probability 1 under medium effort, and 120 with probability 1 under high effort. The set of plausible agent beliefs is the set of beliefs consistent with common knowledge of the relationship between effort and expected output.

Suppose the principal in Example 1 recommends high effort and pays the agent

$3/10 = (20-8)/(120-80) > (20-0)/(120-40)$ of output. The agent's high effort payoff $16 = (3/10) \times 120 - 20$ equals his medium effort payoff $16 = (3/10) \times 80 - 8$ and exceeds his low effort payoff $12 = (3/10) \times 40 - 0$ for every plausible belief; the principal's recommendation is unambiguously incentive compatible; and net guaranteed output is $84 = (7/10) \times 120$. Because this guarantee exceeds both the 40 gross output produced by low effort and the 80 gross output produced by medium effort, optimal contracts must robustly incentivize high effort. As we show in our first theorem, the best contracts for those purposes are linear, and the 3/10 contract is the least generous linear contract under which high effort is incentive compatible. The principal's contract is uniquely optimal.

Second, regarding Theorem 2, consider the following modified version of our first example in which medium effort is less costly for the agent and uniformly incentivizing high effort is therefore more expensive for the principal.

Example 2. The principal's goal is to maximize guaranteed expected net output $y - w(y)$. Output lies in the interval $[0, 160]$ and the agent chooses between low effort at cost 0, medium effort at cost 5, and high effort at cost 20. The principal believes that output is 40 with probability 1 under low effort, 80 with probability 1 under medium effort, and 120 with probability 1 under high effort. The set of plausible agent beliefs is the set of beliefs consistent with common knowledge of the relationship between effort and expected output.

Suppose the principal in Example 2 pays the agent with the linear bonus contract $\max(y-80, 0)$ that returns to him all[2] output in excess of 80 and recommends the agent choose either medium or high effort. Because the agent's piece rate is nondecreasing, his payoff is lowest when he believes output is deterministic and highest when he believes it is distributed in an all-or-nothing manner. Accordingly, his maximal low effort monetary payoff is $20 = (3/4) \times 0 + (1/4) \times (160 - 80)$ and his minimal high effort monetary payoff is $40 = 120 - 80$. While the highest plausible expected wage under medium effort $40 = (1/2) \times 0 + (1/2) \times (160 - 80)$ is such that the agent might strictly prefer medium effort to high, he unambiguously prefers high effort to low.

[2]Linear bonus contracts $\max(\alpha y + \beta, 0)$ with $\alpha = 1$ and $\beta < 0$ are identical to the debt contracts that appear elsewhere in the moral hazard literature (Innes (1990)). The optimality of the $\alpha = 1$ share in Example 2 is coincidental, and debt contracts play no special role in our analysis.

Accordingly, the principal's recommendation is incentive compatible; she believes the agent produces either 80 output at wage 0 or 120 output at wage 40; and her guaranteed payoff 80 is an upper bound for any contract that does not uniformly incentivize high effort. Because the best contract that does uniformly incentivize high effort pays the agent $3/8 = (20-5)/(120-80) > (20-0)/(120-40)$ of output and leaves the principal with only $(5/8) \times 120 = 75$ output in net, the linear bonus contract is optimal.

Finally, regarding Theorem 3, our first two examples suggest and our first two theorems confirm that optimal contracts in our model are simple. At the same time, we have implicitly constrained our principal to the deterministic use of a single contract, and randomizing over contracts is known to improve payoff guarantees in other models of robust contracting with privately informed agents (Kambhampati (2023); Kambhampati, Peng, Tang, Toikka, and Vohra (2025)). While optimal mechanisms in expanded design spaces that allow for randomization, screening, or both, there are straightforward conditions under which our optimal contracts retain their optimality in relaxed formulations of the principal's problem.

Example 3. The principal's goal is to maximize guaranteed expected net output $y - w(y)$. Output lies in the interval $[0, 250]$ and the agent chooses between low effort at cost 0, medium effort at cost 5, and high effort at cost 20. The principal believes that output is 40 with probability 1 under low effort, 80 with probability 1 under medium effort, and 120 with probability 1 under high effort. The set of plausible agent beliefs is the set of beliefs consistent with common knowledge of the relationship between effort and expected output.

Consider the counterfactual formulation of Example 3 in which the agent is known to hold the plausible belief that output is 0 with probability 5/6 and 240 with probability 1/6 under low effort; 0 with probability 1/3, 80 with probability 1/2, and 240 with probability 1/6 under medium effort; and 80 with probability 1/2, 120 with probability 1/3, and 240 with probability 1/6 under high effort. For any contract w, the counterfactual agent's marginal expected wage is $(1/2) \times (w(80) - w(0))$ for increasing low effort to medium and $(1/3) \times (w(120) - w(0))$ for increasing medium effort to high. If medium effort is incentive compatible then $w(80) \geq w(80) - w(0) \geq 2 \times (5-0) = 10$.

Alternatively, if high effort is incentive compatible then $w(120) \geq w(120) - w(0) \geq 3 \times (20-5) = 45$. The principal's payoff in the counterfactual is at most $120 - 45 = 75$; the linear contract that pays the agent 3/8 of output achieves that upper bound and is therefore optimal in both the factual and counterfactual; and the principal does not benefit in worst-case terms from screening or randomizing over contracts.

The optimality of the linear contract in the counterfactual formulation of Example 3 can be understood by analogy to an alternative model. For every contract w, that agent's marginal monetary payoff for increasing low effort to medium effort is as if he were instead paid with the affine contract w^{med} under which (i) the agent's wage for producing 0 output is the same as under w and (ii) the principal's payoff under medium effort is the same as under w. Simultaneously, his marginal monetary payoff for increasing medium effort to high effort is as if he were paid with the potentially distinct affine contract w^{high} under which (i) the agent's wage for producing 0 output is again the same as under w and (ii) the principal's payoff under high effort is the same as under w. In addition to driving our functional form result, this feature suggests a behavioral interpretation of our model as one in which the precise output generating process is common knowledge and the agent mistakenly interprets nonlinear incentive schemes as if they were linear. We formalize this connection and discuss its empirical relevance at the end of Section 3.3.

The paper is organized as follows. In Section 2 we lay out the model and characterize the agent's payoffs. We present our analysis in Sections 3 and 4, review the literature in Section 5, and conclude in Section 6. Omitted proofs are in the Appendix.

2 Model

The agent's choice of *effort* e from set $\mathcal{E}$ determines the distribution of *output* y in set $\mathcal{Y}$. Output is public and received by the principal; effort is private and costly for the agent. The principal provides financial incentives for effort by paying the agent with a *contract* $w : \mathcal{Y} \to \mathbb{R}_+$ under which non-negative wages vary continuously with output. In places, we use the notation t to indicate a generic transfer to the agent.

Formalities The set of efforts $\mathcal{E} \subset \mathbb{R}$ is finite, normalized by $\min \mathcal{E} \equiv 0$, and includes at least two elements; the set of outputs $\mathcal{Y} \subset \mathbb{R}$ is a nondegenerate interval $[\underline{y}, \overline{y}]$. The real numbers $\mathbb{R}$ and their products have the Euclidean topology and finite sets have the discrete topology. We write $\Delta(X)$ for the Borel probabilities on set X and $\delta(x)$ for the degenerate distribution at point x. Given a function $f : \mathcal{Y} \to \mathbb{R}$ and a distribution $G \in \Delta(\mathcal{Y})$, we write $\langle f, G \rangle \equiv \int_{\mathcal{Y}} f(y) \, \mathrm{d}G$ for the expected value of f with respect to G and abusively write $\langle y, G \rangle \equiv \int_{\mathcal{Y}} y \, \mathrm{d}G$ for expected output under G.

Contracts Contract w is an *affine contract* if there exist constants α, β such that $w(y) = \alpha y + \beta$; a *linear contract* if w is an affine contract with minimum payment 0 on $\mathcal{Y}$; and a *linear bonus contract* if there exist constants α, β such that $w(y) = \max(\alpha y + \beta, 0)$ and w has minimum payment 0. By definition, w is linear if and only if it is both affine and a linear bonus contract.

Beliefs The principal and the agent believe output is distributed according to $P : \mathcal{E} \to \Delta(\mathcal{Y})$ and $A : \mathcal{E} \to \Delta(\mathcal{Y})$, respectively. The principal does not know A but does know some non-empty set $\mathcal{A}$ to which it belongs. We call beliefs in $\mathcal{A}$ *plausible* and sometimes refer to the agent's beliefs as his *type.*

We maintain two assumptions about beliefs throughout. First, if (i) A is plausible; (ii) $B(e)$ has either degenerate or binary support for all efforts e; and (iii) $B(e)$ satisfies $\langle y, B(e) \rangle = \langle y, A(e) \rangle$ for all efforts e, then B is also plausible. This richness assumption imposes broad uncertainty about the shape of the agent's beliefs on the principal and is therefore substantive. Second, we assume that the principal's beliefs P and every plausible agent belief A satisfy[3]

$$\underline{y} < \min_e \langle y, A(e) \rangle < \max_e \langle y, A(e) \rangle < \overline{y}, \quad \underline{y} < \min_e \langle y, P(e) \rangle < \max_e \langle y, P(e) \rangle < \overline{y}.$$

In doing so, we rule out extreme situations in which beliefs are necessarily degenerate. Finally, if $P(e)$ has full support for each effort e then we say that *the principal's beliefs have full support.* As we show, linear bonus contracts are uniquely optimal in

[3]In principle, we allow for cases in which beliefs about output under effort e approach $\delta(\underline{y})$ or $\delta(\overline{y})$. This has no bearing on our formal results Lemma 1 and Theorem 5 that treat cases of our problem in which the principal is uncertain about the agent's beliefs about expected output.

specifications of our problem that satisfy this criterion.

Preferences The agent is a risk neutral expected utility maximizer with quasilinear effort costs. The principal is an expected utility maximizer with quasilinear transfer costs and preferences over actions described by utility $\pi : \mathcal{E} \to \mathbb{R}$. Given a contract w and beliefs A, the *agent's problem*

$$\max_{e \in \mathcal{E}} \langle w, A(e) \rangle - e$$

is to maximize his expected utility. In addition to specifying the contract, the principal also *recommends* a deterministic and type-specific tiebreaking rule $c : \mathcal{A} \to \mathcal{E}$ under which $c(A)$ is constrained to be a solution to the agent's problem for every plausible belief A.[4] The *principal's problem*

$$\begin{aligned} &\max_{(w,c)} \min_{A \in \mathcal{A}} \ \pi(c(A)) - \langle w, P(c(A)) \rangle \\ &\text{subject to } \forall A \ \forall e \ \langle w, A(c(A)) \rangle - c(A) \geq \langle w, A(e) \rangle - e \end{aligned}$$

is thus to choose a contract w and a recommendation c to maximize her worst-case payoff against the set of plausible beliefs, noting that the inner minimization problem has a solution because there are finitely many efforts and the principal's payoff does not depend directly on the agent's beliefs. We write $W = (w, c)$ for a typical contract–recommendation pair, $\Pi(W|A) \equiv \pi(c(A)) - \langle w, P(c(A)) \rangle$ for her payoff given pair $W = (w, c)$ and plausible belief A, and $\Pi(W) \equiv \min_{A \in \mathcal{A}} \Pi(W|A)$ for her payoff *guarantee* under W.

The agent's payoffs The agent's payoffs have a simple and useful geometric characterization, as we illustrate in Figure 1. First, we write $\mathscr{A} \equiv \{e \mapsto \langle y, A(e) \rangle | A \in \mathcal{A}\}$ for the set of plausible beliefs about expected output and a for a typical element of that set. Second, given a contract w, we write $\Phi(w) \subset \mathcal{Y} \times \mathbb{R}_+$ for the convex hull of its graph, noting that $\Phi(w)$ is compact because $\mathcal{Y}$ is compact and w is continuous.

[4] We interpret the incentive compatibility constraint in the principal's problem as part of the definition of the recommendation c. We are explicit about this constraint only when introducing that problem and its variants. Otherwise, when we take a contract–recommendation pair $W = (w, c)$ as given, we implicitly assume that c is incentive compatible for w.

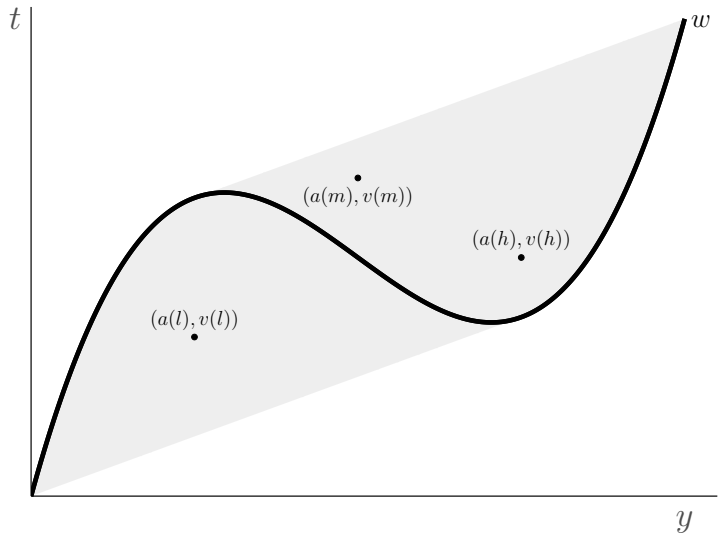


Figure 1: Expected output and wage valuation pairs for a nonlinear contract

Finally, we call a vector $v : \mathcal{E} \to \mathbb{R}_+$ a *wage valuation* for w if there exists a plausible belief A with $\langle w, A(e)\rangle = v(e)$ for all e.

Lemma 1. *Payoff vector v is a wage valuation for contract w if and only if there exists an output vector $a \in \mathscr{A}$ such that $(a(e), v(e)) \in \Phi(w)$ for all efforts e.*

The proof of Lemma 1 follows from the Fenchel-Bunt strengthening of Caratheodory's theorem for connected sets. We make use of it throughout our analysis, both for the purposes of proving our results and also for representing our arguments diagrammatically.

Fact 1. *The interior of payoff set $\Phi(w)$ is empty if and only if w is an affine contract.*

Fact 2. *If output distribution F has full support, then $(\langle y, F\rangle, \langle w, F\rangle)$ belongs to the interior of $\Phi(w)$ if and only if w is not an affine contract.*

Fact 1 implies that affine contracts pin down the agent's payoff up to his beliefs about expected output and is therefore useful for establishing the optimality of linear contracts in the appropriate specifications of our model. Conversely, Fact 2 has implications for tiebreaking and is ultimately useful for establishing some of our uniqueness results. While both results are seemingly standard, we include proofs in the Appendix.

Discussion We make two comments about this setup. First, our agent controls the distribution of output only through his one-dimensional choice of effort.[5] Second, because our principal's preferences over actions are represented by an arbitrary utility function $\pi : \mathcal{E} \to \mathbb{R}$, it might be the case that output is either desirable, undesirable[6], or a pure signal with no tangible value, and there are indeed specifications of our problem in which optimal wages are decreasing in output. While our setup is general enough to include the specification $\pi(e) = \langle y, P(e) \rangle$ under which the principal's objective is to maximize net expected output, we impose that as an assumption only in Theorem 5. Outside of that context, our output variable is best understood as a signal of the agent's action rather than as a direct determinant of the principal's payoff.

These two features of our model lie in contrast to the robust contracting framework developed in Carroll (2015), in which (i) the agent controls both effort and the distribution of output and (ii) the principal's objective is to maximize net expected output. There, linear contracts are optimal because they align the risk-neutral agent's incentives for risk taking with the principal's risk neutral preferences. Here, linear bonus contracts are optimal because they balance the tradeoff between providing the right incentives for effort and paying the agent economical wages. Because our results do not in general require that the principal's objective is to maximize net output, our optimal contracts do not necessarily achieve the same preference alignment that they do in Carroll's setting, even where they are fully linear.

3 Common knowledge of the relationship between effort and expected output

The model described in Section 2 is permissive. While we allow for situations in which the principal knows the agent's beliefs about expected output, we also allow for situations in which she views every belief that is consistent with our interiority

[5]All of our results go through without modification if we allow the agent's technology to be multidimensional in the sense that distinct actions share effort costs.

[6]We imagine an employer (professional sports league) who disincentivizes flagrant rule violations via fines, a parent who punishes bad behavior with deductions from a child's allowance, or a government that pays a company to reduce its production of negative externalities (pollution).

criteria as plausible. The principal's problem is most interesting to study when she is willing to rule out extreme heterogeneity within the set of plausible beliefs, and we focus first on its simplest structured specification.

Assumption 1. *(Common knowledge of the relationship between effort and expected output) If beliefs A are plausible then $\langle y, A(e)\rangle = \langle y, P(e)\rangle$ for every effort e.*

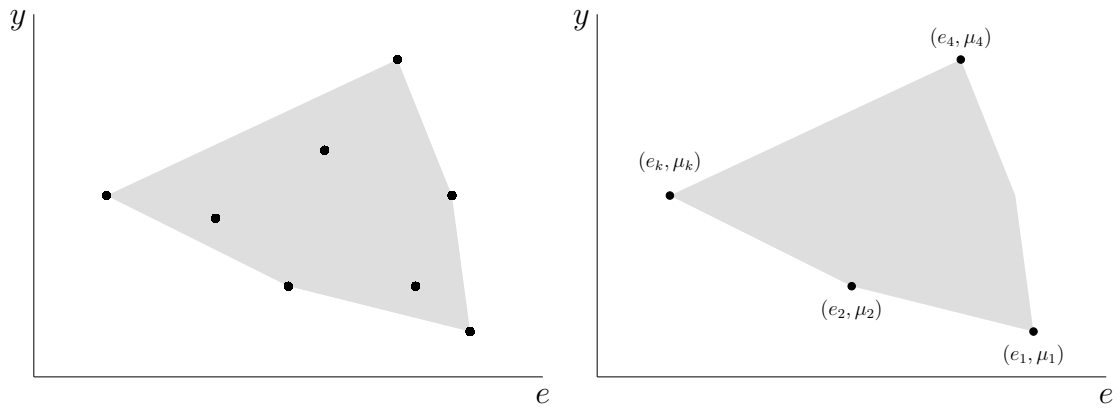
We maintain Assumption 1 throughout this section of the paper and write μ for the mutually understood expected output map $e \mapsto \langle y, P(e)\rangle$. Our analysis of the model with common knowledge of the relationship between effort and expected output is organized as follows. First, in Section 3.1 we study a constrained version of the principal's problem in which she recommends the same level of effort to every type of agent. Next, in Section 3.2, we treat the principal's problem in full. Finally, in Section 3.3, we consider expanded versions of the principal's problem and identify broad conditions under which she does not benefit from screening the agent, randomizing over contracts, or employing other types of complex mechanisms. The first subsection supports the third.

3.1 The uniform effort problem

Our principal's preference for robustness suggests that she might do well to pay the agent with a contract under which there is certainty about his actions. Accordingly, we initiate our study of the principal's problem by first considering the *uniform effort problem*

$$\max_{(w,c)} \min_{A \in \mathcal{A}} \pi(c(A)) - \langle w, P(c(A))\rangle$$
$$\text{subject to } \forall A\ \forall e'\ \langle w, A(c(A))\rangle - c(A) \geq \langle w, A(e')\rangle - e' \text{ and } \forall A\ c(A) = e$$

in which the principal constrains herself to incentive schemes that assign the same effort e to every type of agent. As we show in the major result of this section (Theorem 1), the best contracts for incentivizing uniform effort are linear contracts. After establishing that result, we show how it leads to the optimality of linear contracts in two-effort specifications of the principal's problem.



(a) efforts (b) efficient efforts

Figure 2: Identifying efficient actions

The first step in our analysis of the uniform effort problem is to identify the set of efforts for which that problem is feasible. While specifications of our model that satisfy the common knowledge of expected output assumption are sufficiently well structured that the principal's problem is nontrivial and interesting to study, it might still be the case that the relationship between effort and expected output μ is nonmonotone, nonconvex, noninjective, or pathological in other ways.

Definition 1. Effort e is *efficient* if $e \leq \varepsilon$ for every ε such that $(\varepsilon, \mu(e))$ belongs to the convex hull of graph(μ).

The set of efficient actions is the set of actions under which effort costs are convex in expected output, as we illustrate in Figure 2. We enumerate the efficient efforts as $e_1, ..., e_m$ so that the corresponding sequence of expected outputs $\mu_i \equiv \mu(e_i)$ is strictly increasing in i. Further, we write k for the index i associated with the necessarily-efficient 0 effort action and $P_i \equiv P(e_i), \pi_i \equiv \pi(e_i)$ for the principal's beliefs and utility, respectively. It will be helpful to keep in mind that e_i is strictly decreasing in i for $i = 1, ..., k$ and strictly increasing in i for $i = k, ..., m$, and further that $\mu_1 = \min_e \mu(e), \mu_m = \max_e \mu(e)$. Accordingly, if expected output is strictly increasing in effort then $k = 1$ and e_i is strictly increasing in i. The reader uninterested in generality might wish to restrict attention to those cases.

Our efficiency criterion begins to suggest a special role to be played by linear contracts because it depends only on effort and expected output. In order to state our

formal results, it will be useful to first identify the linear contract that most cheaply incentivizes each efficient effort. For each i, define

$$w_i(y) \equiv \begin{cases} \dfrac{e_i - e_{i+1}}{\mu_{i+1} - \mu_i}(\overline{y} - y) & i < k, \\ 0 & i = k, \\ \dfrac{e_i - e_{i-1}}{\mu_i - \mu_{i-1}}(y - \underline{y}) & i > k; \end{cases} \quad t_i \equiv \begin{cases} \dfrac{e_i - e_{i+1}}{\mu_{i+1} - \mu_i}(\overline{y} - \mu_i) & i < k, \\ 0 & i = k, \\ \dfrac{e_i - e_{i-1}}{\mu_i - \mu_{i-1}}(\mu_i - \underline{y}) & i > k; \end{cases} \quad \Pi_i \equiv \pi_i - t_i,$$

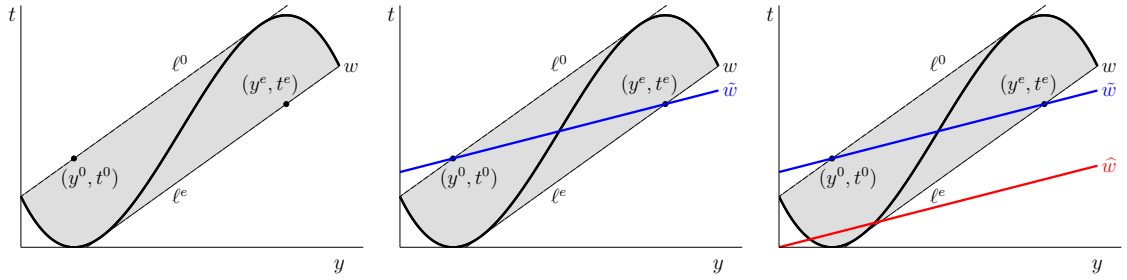
noting that the share parameter for contract w_i is derived from the binding incentive compatibility constraint associated with reducing effort e_i to the marginally-lower efficient effort e_{i-1} or e_{i+1}, as appropriate. Let recommendation c_i satisfy $c_i(A) \equiv e_i$ for all plausible beliefs A and set $W_i \equiv (w_i, c_i)$. Our first major result confirms that the uniform effort problem is feasible for the efficient efforts and that the best uniform effort contracts are linear.

Theorem 1. *Let Assumption 1 hold. First, the uniform effort problem is feasible for effort e if and only if $e = e_i$ is efficient. Second, W_i solves the uniform effort problem for effort e_i with value $\Pi(W_i) = \Pi_i$. Third, if e_i is nonzero or if the principal's beliefs have full support, then W solves the uniform effort problem for effort e_i only if $W = W_i$.*

We prove Theorem 1 in three steps, each of which we present as a standalone lemma. First, we confirm that efficiency characterizes feasibility.

Lemma 2. *Let Assumption 1 hold. The uniform effort problem is feasible for effort e if and only if $e = e_i$ is efficient.*

The sufficiency of efficiency follows straightforwardly from the separating hyperplane theorem, which yields uniform effort linear contracts for those actions. Conversely, we establish necessity by constructing plausible beliefs under which the agent's monetary payoff necessarily varies linearly with the expected output that he produces, regardless of how he is paid. As suggested by our definition, inefficient efforts are never incentive compatible when the agent's payoffs are a linear function of expected output.



(a) original contract (b) weakly improved contract (c) strictly improved contract

Figure 3: Constructing an improving linear contract

In the second step, we show that any nonlinear contract that uniformly incentivizes a particular level of effort can be improved to a linear contract that does the same. This is the heart of the argument, and we present the proof in the body.

Lemma 3. *Let Assumption 1 hold. If $W = (w, c)$ is feasible for the uniform effort problem with effort e and w is not a linear contract, then there exists a linear contract $\widehat{w}$ such that $\widehat{W} \equiv (\widehat{w}, c)$ is feasible for the uniform effort problem with effort e and satisfies $\Pi(\widehat{W}) \geq \Pi(W)$, with $\Pi(\widehat{W}) > \Pi(W)$ if e is nonzero or if the principal's beliefs have full support.*

The intuition for Lemma 3 is mostly captured by Fact 1, where we remarked that there is certainty about the agent's payoffs if and only if the principal pays the agent with an affine contract. Nonlinearities in the agent's wages yield beliefs under which the agent's payoff is relatively low for the intended level of effort and relatively high for every alternative level of effort, vis-à-vis the principal's own beliefs about transfers. Those asymmetries manifest as an endogenous "complexity penalty" that must be paid by the principal in order to uniformly incentivize the intended effort, and the magnitude of that penalty is such that it is less expensive for her to simply pay the agent with a linear contract if she desires certainty about his actions.

Proof of Lemma 3. Suppose $W = (w, c)$ is feasible for the uniform effort problem with effort e and that w is not a linear contract. If w is affine, set $\widehat{w} \equiv w - \min_z w(z)$. Alternatively, if e is 0, set $\widehat{w} \equiv w_k$. In both cases, $\widehat{W} \equiv (\widehat{w}, c)$ yields the claimed improvement. If instead w is not affine and e is nonzero, write $y^0 \equiv \mu(0), y^e \equiv \mu(e), t^0 \equiv \max\{t|(y^0, t) \in \Phi(w)\}, t^e \equiv \min\{t|(y^e, t) \in \Phi(w)\}$ and note that $t^e > t^0$ per

our hypothesis that W is feasible for the uniform effort problem with effort e.

The supporting hyperplane theorem provides a support line $\ell^0 \subset \mathbb{R}^2$ to $\Phi(w)$ at (y^0, t^0) and a support line $\ell^e \subset \mathbb{R}^2$ to $\Phi(w)$ at (y^e, t^e). First, because y^0, y^e are interior, ℓ^0, ℓ^e are nonvertical and can therefore be interpreted as affine functions $\ell^0, \ell^e : \mathbb{R} \to \mathbb{R}$. Second, because w is not affine, Fact 2 implies

$$\begin{aligned} \max\{t|(y^0, t) \in \Phi(w)\} &> \min\{t|(y^0, t) \in \Phi(w)\}, \\ \max\{t|(y^e, t) \in \Phi(w)\} &> \min\{t|(y^e, t) \in \Phi(w)\}. \end{aligned} \tag{1}$$

In turn, because $\ell^0(y^0) = \max\{t|(y^0, t) \in \Phi(w)\}$ and $\ell^e(y^e) = \min\{t|(y^e, t) \in \Phi(w)\}$, we obtain for all pairs $(y, t) \in \Phi(w)$

$$\ell^0(y) \geq t \geq \ell^e(y). \tag{2}$$

Consider the affine contract[7] $\tilde{w}$ with $\tilde{w}(y^0) \equiv t^0$ and $\tilde{w}(y^e) \equiv t^e$, as depicted in Figure 3. Jointly from (1)–(2) and our choice of $\tilde{w}$, we have $\ell^e(y^0) < \ell^0(y^0) = \tilde{w}(y^0) = t^0 < t^e = \tilde{w}(y^e) = \ell^e(y^e) < \ell^0(y^e)$. In turn, because $\ell^0, \ell^e, \tilde{w}$ are affine, we have (i) $\tilde{w}(y) < t^0 \iff \tilde{w}(y) > \ell^0(y)$ and (ii) $\tilde{w}(y) > t^e \iff \tilde{w}(y) < \ell^e(y)$. Further, because $\Phi(w)$ is convex by definition and $(y^0, \tilde{w}(y^0) = t^0), (y^e, \tilde{w}(y^e) = t^e) \in \Phi(w)$ per our choice of t^0, t^e, we also have

$$t^0 \leq \tilde{w}(y) \leq t^e \iff (y, \tilde{w}(y)) \in \Phi(w). \tag{3}$$

Together, (2)–(3) have two implications of significance. First, if $\tilde{w}(y) > \max\{t|(y, t) \in \Phi(w)\}$ then $\tilde{w}(y) < t^0$. Because $\tilde{w}(y^e) = t^e$ by construction and $t^e - e \geq t^0 - 0$ by hypothesis, efforts that yield wages strictly smaller than t^0 are not incentive compatible under contract $\tilde{w}$. Accordingly, $\tilde{W} \equiv (\tilde{w}, c)$ uniformly incentivizes effort e. Second, if $\tilde{w}(y) < t^0$ then $\tilde{w}(y) > \max\{t|(y, t) \in \Phi(w)\}$. Because y^0 is interior, $\tilde{w}$ is affine, and $\tilde{w}(y^0) = t^0$, the minimum of $\tilde{w}$ is strictly less than t^0. In turn, because the graph of w is contained in $\Phi(w)$ by definition, the minimum of $\tilde{w}$ is strictly greater than the minimum of w. This implies both that $\tilde{w}$ is a well defined contract and also that the linear contract $\hat{w} \equiv \tilde{w} - \min_z \tilde{w}(z)$ satisfies

[7]We show near the end of the proof that $\tilde{w}$ satisfies limited liability and is therefore a contract.

$\langle \widehat{w}, P(e) \rangle < \langle \tilde{w}, P(e) \rangle = \min\{t | (y^e, t) \in \Phi(w)\rangle \leq \langle w, P(e) \rangle$. The pair $\widehat{W} \equiv (\widehat{w}, c)$ inherits feasibility for the uniform effort problem with effort e from $\tilde{W}$ and satisfies $\Pi(\widehat{W}) > \Pi(\tilde{W}) \geq \Pi(W)$. □

Lemma 3 implies that optimal uniform effort contracts are linear. The third and final step of the proof of Theorem 1 is to verify that W_i is optimal among the set of incentive schemes that are feasible for the uniform effort problem with effort e_i.

Lemma 4. *Let Assumption 1 hold. First, if $W = (w, c)$ is feasible for the uniform effort problem with effort e_i and w is a linear contract, then W solves the uniform effort problem for effort e_i if and only if $W = W_i$. Second, the uniform effort problem for effort e_i has value $\Pi(W_i) = \Pi_i$.*

The proof of Lemma 4 is a straightforward accounting exercise, and we leave the details to the Appendix. With that result in hand, we obtain Theorem 1 as a summary of Lemmas 2–4.

In order to understand the implications of our first theorem for optimality in the principal's problem, recall that we began our discussion of Example 1 by arguing that the optimal contract must incentivize high effort in every type of agent. Once that has been established, it is enough to show that the best uniformly incentive compatible contract is linear, as Lemma 3 implies. However, that the optimal contract must robustly incentive high effort is a specific feature of that example rather than a general feature of our model, as Example 2 makes clear. Instead, the results of this section are only strong enough to guarantee the linearity of solutions to the principal's problem if there are two levels of effort and the principal's beliefs have full support.

Proposition 1. *Let Assumption 1 hold and suppose there are exactly two levels of effort and the principal's beliefs have full support. Contract–recommendation pair W solves the principal's problem if and only if there exists an index i such that $W = W_i$ and $\Pi_i = \max(\Pi_1, \Pi_2)$.*

If there are two actions available to the agent and the principal's beliefs have full support, any contract that pays positive wages for any output either uniformly incentivizes the more costly of the two actions or yields a strictly worse guarantee than the trivial contract that pays the agent wage 0 for every output. Once that has been verified, Proposition 1 follows immediately from Theorem 1.

3.2 The principal's problem

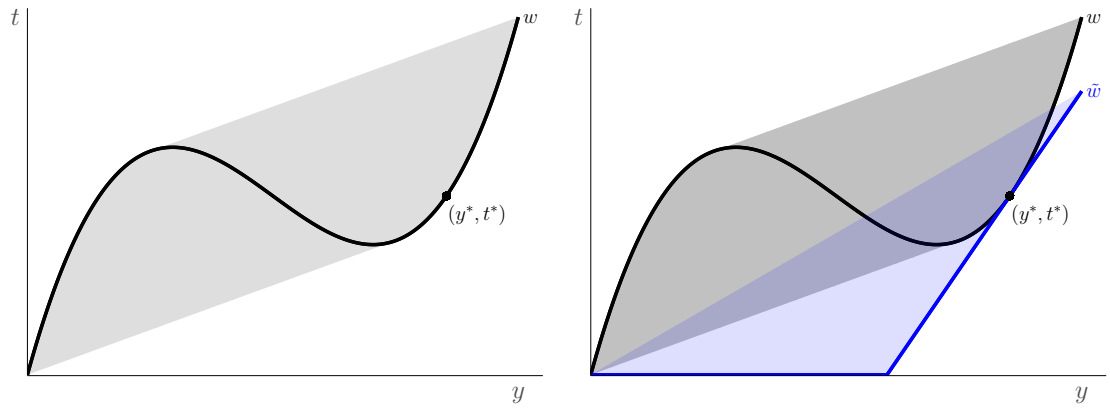
The binary effort assumption in Proposition 1 is restrictive. At the same time, the uniform effort constraint is arguably unnatural and, as we have shown in Example 2, at least sometimes binding. In this section, we show that linear bonus contracts are optimal for general specifications of the principal's problem, and uniquely so if her beliefs have full support.

Theorem 2. *Let Assumption 1 hold. First, there exists a linear bonus contract w and a recommendation c such that $W = (w, c)$ solves the principal's problem. Second, if the principal's beliefs have full support, then $W = (w, c)$ solves the principal's problem only if w is a linear bonus contract.*

The substance of Theorem 2 is contained in the following lemma.

Lemma 5. *Let Assumption 1 hold. If $W = (w, c)$ and w is not a linear bonus contract, then there exists a linear bonus contract $\widehat{w}$ and a recommendation $\widehat{c}$ such that $\widehat{W} \equiv (\widehat{w}, \widehat{c})$ satisfies $\Pi(\widehat{W}) \geq \Pi(W)$, with $\Pi(\widehat{W}) > \Pi(W)$ if the principal's beliefs have full support.*

Lemma 5 can be understood by careful consideration of the geometric structure of the agent's payoff set, as described by Lemma 1. For the purposes of illustration, let w be any contract, c any incentive compatible recommendation, A^* any plausible belief that satisfies $\langle w, A(e)\rangle = \min_{B\in\mathcal{A}}\langle w, B(e)\rangle$ for each effort e, and e^* the level of effort assigned to the minimal-payoff agent A^*. As we show in the proof, there exists a corresponding linear bonus contract $\widehat{w}$ that preserves the agent's lowest-possible payoff for effort e^* despite paying the agent no more than w for any level of output. When coupled with any incentive compatible recommendation $\widehat{c}$ that breaks ties in favor of effort e^* wherever possible, the improved contract $\widehat{w}$ not only directly improves the principal's payoff by unambiguously reducing wages but also makes efforts other than e^* less attractive to the agent, regardless of his beliefs. The set of actions the agent might take under the improved contract is a subset of the actions he might take under the original contract, and her payoff for each of those actions is at least as high. In light of the principal's preferences for robustness, this improves her payoff guarantee.



(a) original contract w (b) improved contract $\tilde{w}$

Figure 4: Constructing an improving linear bonus contract

Proof of Lemma 5. Suppose w is not a linear bonus contract, let c be any incentive compatible recommendation, and set $W \equiv (w, c)$. As in the proof of Theorem 1, if w is an affine contract then (i) $\widehat{w} \equiv w - \min_z w(z)$ is a linear bonus contract and (ii) $\widehat{W} \equiv (\widehat{w}, c)$ satisfies $\Pi(\widehat{W}) > \Pi(w)$. We devote the rest of the proof to the more interesting cases in which w is not an affine contract. Let plausible belief A^* satisfy $\langle w, A^*(e)\rangle = \min\{t|(\mu(e), t) \in \Phi(w)\}$ for all efforts e and write $e^* \equiv c(A^*), y^* \equiv \mu(e^*), t^* \equiv \langle w, A^*(e^*)\rangle$. Per our choice of A^*, the pair (y^*, t^*) lies on the boundary of $\Phi(w)$. In turn, the supporting hyperplane theorem provides a normal vector $\eta = (\eta_y, \eta_t) \neq (0, 0)$ such that

$$\forall (y, t) \in \Phi(w)\ \eta \cdot (y^*, t^*) \geq \eta \cdot (y, t), \tag{4}$$

where (i) η_t is nonzero because y^* is interior and (ii) η_t is nonpositive because $\Phi(w)$ contains points (y^*, t) with $t > t^*$ per Fact 2. Jointly, the support line $\{(y, t)|\eta \cdot (y, t) = \eta \cdot (y^*, t^*)\}$ and the limited liability constraint delineate a contract $\tilde{w}(y) \equiv \max(t^* + (\eta_t)^{-1}\eta_y(y^* - y), 0)$, as depicted in Figure 4. We make two observations about $\tilde{w}$. First, because the graph of w is contained in $\Phi(w)$ by definition, $\eta \cdot (y^*, t^*) \geq \eta \cdot (y, w(y))$ for all y per (4). In turn, limited liability and $\eta_t < 0$ imply

$$w(y) \geq \max(t^* + \frac{\eta_y}{\eta_t}(y^* - y), 0) = \tilde{w}(y). \tag{5}$$

Second, because $\tilde{w}$ is the pointwise maximum of two convex functions, $\tilde{w}$ is itself convex. In turn, Jensen's inequality implies

$$\min\{t|(y^*,t) \in \Phi(\tilde{w})\} = \tilde{w}(y^*) = t^* = \min\{t|(y^*,t) \in \Phi(w)\}. \tag{6}$$

Let $\tilde{c}$ be any incentive compatible tiebreaking rule for contract $\tilde{w}$ that breaks ties in favor of e^* whenever e^* is a solution to the agent's problem and consider $\tilde{W} \equiv (\tilde{w}, \tilde{c})$. While the condition

$$\max\{t|(\mu(e),t) \in \Phi(\tilde{w})\} - e > \min\{t|(y^*,t) \in \Phi(\tilde{w})\} - e^* \tag{7}$$

is necessary for the existence of a plausible belief A with $\tilde{c}(A) = e$, the condition

$$\max\{t|(\mu(e),t) \in \Phi(w)\} - e > \min\{t|(y^*,t) \in \Phi(w)\} - e^* \tag{8}$$

is sufficient for the existence of a plausible belief A with $c(A) = e$. At the same time, (5) and (6) imply that (7) is sufficient for (8). Accordingly, $\{\tilde{c}(A)|A \in \mathcal{A}\} \subset \{c(A)|A \in \mathcal{A}\}$. Furthermore, (5) implies $\pi(e) - \langle \tilde{w}, P(e)\rangle \geq \pi(e) - \langle w, P(e)\rangle$ for all efforts e, with strict inequality if the principal's beliefs have full support. Accordingly, we have $\Pi(\tilde{W}) \geq \Pi(W)$, with strict inequality again if the principal's beliefs have full support. If $\tilde{w}$ has minimum payment 0 then $\tilde{w}$ is a linear bonus contract and the proof is complete. If $\tilde{w}$ does not have minimum payment 0, then $\widehat{w}(y) \equiv \tilde{w}(y) - \min_z \tilde{w}(z)$ is a linear bonus contract and $\widehat{W} \equiv (\widehat{w}, \tilde{c})$ satisfies $\Pi(\widehat{W}) > \Pi(\tilde{W}) \geq \Pi(W)$. □

Theorem 2 follows jointly from Lemma 5 and the existence of an optimal contract. We provide a detailed proof of the latter in the Appendix.

3.3 The principal's expanded problem

Thus far, we have constrained the principal to deterministically pay every type of agent with the same contract. At the same time, her aversion to uncertainty is such that she would be guaranteed to at least weakly benefit from learning the agent's beliefs before settling on an incentive scheme. Were that benefit guaranteed to be strict, then it might be possible for her to improve on the best-available payoff guarantee by

screening the agent, randomizing over contracts, or both. However, if instead

$$\max_{W} \min_{A \in \mathcal{A}} \Pi(W|A) = \min_{A \in \mathcal{A}} \max_{W} \Pi(W|A)$$

then the principal's problem has a *saddle point* (W, A) with the property that

$$\forall W' \;\; \Pi(W|A) \geq \Pi(W'|A), \qquad\qquad \forall A' \;\; \Pi(W|A) \leq \Pi(W|A').$$

In turn, (i) W solves the principal's problem; (ii) the principal's payoff guarantee under W is her payoff $\Pi(W|A)$ when the agent's beliefs are A; and (iii) W maximizes the principal's payoff in the counterfactual in which the principal knows the agent's beliefs are A. In all such cases, the optimality of W persists in expanded formulations of the principal's problem that allow for mechanisms that are more complex than the deterministic use of a single contract.

In this section, we identify broad conditions under which the principal's problem indeed has a saddle point and optimal contracts are linear. We then use this result to formalize our observation that the principal does not benefit in worst-case terms from using complex mechanisms. In doing so, we provide a theoretical justification for the deterministic use of a single contract with a simple functional form. We view these as positive results.

Constructing the saddle point

Assumption 2. *(Broad support) First, every belief A with $\langle y, A(e) \rangle = \mu(e)$ for every effort e is plausible. Second, the set of outputs $\mathcal{Y} = [\underline{y}, \overline{y}]$ satisfies*

$$\frac{\mu_1 - \underline{y}}{\mu_m - \mu_1} \geq \frac{1 + \sqrt{5}}{2}, \qquad\qquad \frac{\overline{y} - \mu_m}{\mu_m - \mu_1} \geq \frac{1 + \sqrt{5}}{2}.$$

The second part of Assumption 2 implies that the range of outputs $\overline{y} - \underline{y}$ is roughly four times the length of the range of expected outputs $\mu_m - \mu_1$ or greater. Along with Assumption 1, we maintain it throughout this subsection. Our arguments are

constructive. First, define weights τ_i and remainder r by

$$\tau_i \equiv \begin{cases} \dfrac{\mu_{i+1}-\mu_i}{\overline{y}-\mu_i} & i<k, \\ 0 & i=k, \\ \dfrac{\mu_i-\mu_{i-1}}{\mu_i-\underline{y}} & i>k; \end{cases} \qquad r \equiv \frac{\mu_k - \overline{y}\sum_{i<k}\tau_i - \underline{y}\sum_{i>k}\tau_i}{1-\sum_i \tau_i}.$$

Next, for each efficient effort e_i define belief S by

$$S(e_i) \equiv S_i \equiv \begin{cases} \sum_{j<i}\tau_j\delta(\overline{y}) + \sum_{j=i}^{k}\tau_j P_j + \sum_{j>k}\tau_j\delta(\underline{y}) + (1-\sum_j \tau_j)\delta(r) & i \le k, \\ \sum_{j<k}\tau_j\delta(\overline{y}) + \sum_{j=k}^{i}\tau_j P_j + \sum_{j>i}\tau_j\delta(\underline{y}) + (1-\sum_j \tau_j)\delta(r) & i \ge k. \end{cases}$$

Finally, for each inefficient effort e there exists a distribution $\gamma^e \in \Delta(\{1,..,m\})$ with $\sum_i \gamma_i^e \mu_i = \mu(e)$ and $\sum_i \gamma_i^e e_i < e$. Complete the construction by assigning $S(e) \equiv \sum_i \gamma_i^e S_i$ for each inefficient effort e.

Theorem 3. *Let Assumptions 1–2 hold. First, the principal's problem has a saddle point. Second, (W_i, S) is a saddle point if and only if $\Pi_i = \max(\Pi_1, ..., \Pi_m)$. Third, if the principal's beliefs have full support, then W solves the principal's problem only if $W = W_i$ for some index i with $\Pi_i = \max(\Pi_1, ..., \Pi_m)$.*

In addition to identifying a saddle point and proving that linear contracts are optimal, Theorem 3 explicitly characterizes the set of solutions to the principal's problem. There are three steps to the proof. First, in Lemma 6 we use the underestimates

$$\sum_{i<k}\tau_i \lessapprox \ln\left(\frac{\overline{y}-\mu_1}{\overline{y}-\mu_k}\right) \lessapprox \frac{\mu_k-\mu_1}{\overline{y}-\mu_k}, \qquad \sum_{i>k}\tau_i \lessapprox \ln\left(\frac{\mu_m-\underline{y}}{\mu_k-\underline{y}}\right) \lessapprox \frac{\mu_m-\mu_k}{\mu_k-\underline{y}}$$

to show that S is both well defined and plausible whenever Assumption 2 is satisfied.[8]

Lemma 6. *Let Assumptions 1–2 hold. Belief S is both well defined and plausible.*

[8]It can be readily verified that S is well defined when there are two efforts, regardless of whether or not the second criterion in Assumption 2 is satisfied. Accordingly, Theorem 3 subsumes Proposition 1 in two-effort specifications of our problem that satisfy that Assumption's first criterion.

Second, we show that for any contract w (i) belief S is assigned an efficient effort e_i and (ii) expected transfers to the agent under e_i match or exceed expected transfers under the linear contract w_i that uniformly incentivizes effort e_i. This fact falls straightforwardly out of the incentive compatibility conditions for S and is the crux of the argument.

Lemma 7. *Let Assumptions 1–2 hold. For every $W = (w, c)$ there exists an index i such that $c(S) = e_i$ and $\Pi_i \geq \Pi(W|S)$.*

Proof. Let $W = (w, c)$ be any contract–recommendation pair and note that $\langle w, S(e)\rangle - e = \sum_i \gamma_i^e \langle w, S_i\rangle - e < \sum_i \gamma_i^e (\langle w, S_i\rangle - e_i)$ for all inefficient efforts e. Accordingly, $c(S) = e_i$ for some index i. The lower bounds

$$\begin{aligned} \langle w, S_i\rangle - \langle w, S_{i+1}\rangle &= \tau_i(\langle w, P_i\rangle - w(\overline{y})) \geq e_i - e_{i+1} && i < k, \\ \langle w, S_i\rangle - \langle w, S_{i-1}\rangle &= \tau_i(\langle w, P_i\rangle - w(\underline{y})) \geq e_i - e_{i-1} && i > k \end{aligned}$$

follow directly from the definitions of S_i and the adjacent distributions S_{i-1}, S_{i+1}. We have in turn

$$\begin{aligned} \langle w, P_i\rangle \geq \langle w, P_i\rangle - w(\overline{y}) \geq \frac{e_i - e_{i+1}}{\mu_{i+1} - \mu_i}(\overline{y} - \mu_i) = t_i && i < k, \\ \langle w, P_i\rangle \geq 0 = t_i && i = k, \\ \langle w, P_i\rangle \geq \langle w, P_i\rangle - w(\underline{y}) \geq \frac{e_i - e_{i-1}}{\mu_i - \mu_{i-1}}(\mu_i - \underline{y}) = t_i && i > k, \end{aligned} \tag{9}$$

where the first column of inequalities follows from limited liability and the second from the definition of τ_i. Accordingly, $\Pi(W|S) = \pi_i - \langle w, P_i\rangle \leq \pi_i - t_i = \Pi_i$. □

With support from Theorem 1, Lemma 7 yields a saddle point for the principal's problem. The third and final step of the argument is to show that if the principal's beliefs have full support then nonlinear contracts are suboptimal.

Lemma 8. *Let Assumptions 1–2 hold and suppose the principal's beliefs have full support. If $W = (w, c)$ and w is not a linear contract then there exists an index i such that $\Pi_i > \Pi(W)$.*

Lemma 8 follows jointly from Fact 2 and an adaptation of the argument given in proof of Lemma 7. For the purposes of illustration, suppose there are two efforts $e_1 < e_2$ with $\mu_1 < \mu_2$ and let the principal pay the agent with a contract–recommendation pair $W = (w, c)$ under which w is not a linear contract. If $c(S) = e_1$ then W provides a strictly worse guarantee than W_1 because the principal's beliefs have full support. Otherwise, if $c(S) = e_2$ then there are two cases to consider. First, if $\langle w, S_2\rangle - e_2 > \langle w, S_1\rangle - e_1$ then the incentive compatibility conditions for S imply directly that $\langle w, P_2\rangle > \langle w_2, P_2\rangle$ and hence $\Pi(W_2) > \Pi(W)$. Second, if $\langle w, S_2\rangle - e_2 = \langle w, S_2\rangle - e_2$ then (i) (9) implies w is not an affine contract and (ii) Fact 2 implies there exists a belief type $\tilde{S}$ nearby to S satisfying $\langle w, \tilde{S}_1\rangle - e_1 > \langle w, \tilde{S}_2\rangle - e_2$, noting that S_1, S_2 inherit full support from our hypothesis that the principal's beliefs themselves have full support. In that case, we again have $\Pi(W_1) > \Pi(W)$. With the formal result in hand, we obtain Theorem 3 as an easy consequence of Lemmas 6–8 and the results of Section 3.1. We provide a short proof in the Appendix.

As a final technical matter, the sufficiency of Assumption 2 for the existence of a saddle point in the principal's problem raises a question as to whether or not that problem might have saddle points in specifications that do not satisfy the criterion. While the suboptimality of linear contracts in Example 2 confirms that the principal's problem does not in general have a saddle point in which the constituent optimal contract is linear, it might still be the case that there exists a belief that is adversarial with respect to the linear bonus contract that is optimal in that specification of the problem. Toward a general result, Sion's minimax theorem implies that strong duality holds in our problem if (i) the map $W \mapsto \Pi(W|A)$ is upper semicontinuous and quasiconcave for every plausible A and (ii) the map $A \mapsto \Pi(W|A)$ is lower semicontinuous and quasiconvex for every well defined W. In Appendix D, we construct a four-action example in which both criteria in (ii) are violated. In doing so, we demonstrate that Sion's theorem can not be used to establish the existence of a saddle point in general specifications of our problem.

The principal's expanded problem Among its other implications, Theorem 3 implies that linear contracts remain optimal in expanded formulations of the principal's problem that allow for screening and randomization. Formally, a *menu* $(\mathcal{W}, c)$ is an

incentive compatible assignment of contract $\mathcal{W}(A)$ and action $c(A)$ to each plausible belief; a *random menu* $\nu = (N; (\mathcal{W}_1, c_1), ..., (\mathcal{W}_N, c_N); (\nu_1, ..., \nu_N))$ is a randomization over finitely many menus $(\mathcal{W}_1, c_1), ..., (\mathcal{W}_N, c_N)$ according to probability distribution $(\nu_1, ..., \nu_N)$; the principal's payoff given random menu ν and belief A is $\Pi(\nu|A) \equiv \int \Pi((\mathcal{W}(A), c(A))|A)\, \mathrm{d}\nu$; and the *principal's expanded problem*

$$\max_{\nu=(N;(\mathcal{W}_1,c_1),...,(\mathcal{W}_N,c_N);(\nu_1,...,\nu_N))} \min_{A\in\mathcal{A}} \Pi(\nu|A)$$
$$\text{subject to } \forall j\ \forall A\ \forall B\ \forall e\ \langle \mathcal{W}_j(A), A(c_j(A))\rangle - c_j(A) \geq \langle \mathcal{W}_j(B), A(e)\rangle - e$$

is to choose the random menu that maximizes her guaranteed payoff against the set of plausible beliefs.[9]

Proposition 2. *Let Assumptions 1–2 hold. If* $\Pi_i = \max(\Pi_1, ..., \Pi_m)$ *then* W_i *solves the principal's expanded problem.*

Proposition 2 formalizes our assertion that Theorem 3 provides a worst-case justification for the use of linear contracts in expanded design spaces. While the uniqueness result in that theorem implies our optimal contracts are not weakly dominated by other contracts, it falls short of ruling out weak dominance by random menus. We leave the analysis of that issue for future work.[10]

Discussion Economists have documented extensive evidence that a critical mass of real-world decision makers seem to mistakenly interpret nonlinear incentive schemes as if they were linear (see Ito (2014) for an overview of the evidence and Martimort and Stole (2020) for a nonlinear pricing exercise that takes this behavior as its premise). As Ito shows, these mistakes potentially have negative welfare implications not only for the decision maker but also for the firms with whom they interact. Along those lines, there is a formal sense in which the saddle point agent types in our model themselves

[9]We identify the degenerate random menu $(1; (\mathcal{W}, c); (1))$ satisfying $\mathcal{W}(A) = w$ for all plausible beliefs A with the contract–recommendation pair (w, c).

[10]To see why this is a potential concern, return to Example 3. Screening the agent by giving him a choice between the robustly optimal contract $(3/8)y$ and the linear bonus contract $\max((3/4)y - 90, 0)$ weakly dominates paying him with the former because there are agent types that are indifferent between the two contracts, even though the latter pays no more than the former for every output in the support of the principal's beliefs.

interpret every contract as if it were linear, and this perception drives the optimality of linear contracts in the corresponding specifications of our model. To the extent that our agent's behavior reasonably approximates the real-world behavior documented by Ito (2014) and other authors, our analysis might be viewed as providing a formal model of those phenomena that fits within the standard rational choice framework. From that perspective, this paper (i) justifies the provision of linear incentives in contexts where misunderstandings by the agent are a first order concern and (ii) provides an interpretation of linear contracts as being "simple" because they are easy for the agent to understand, rather than because they are more literally robust to various forms of uncertainty.

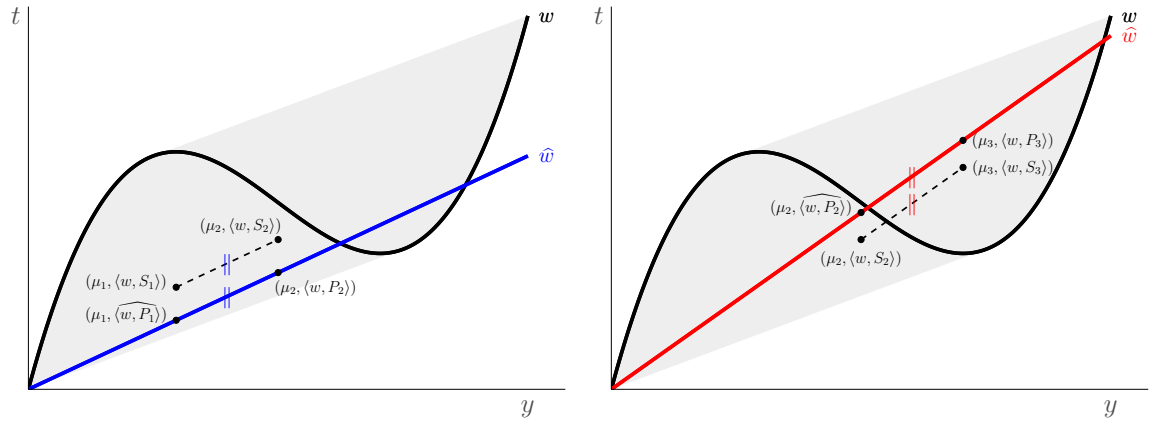
To make the connection concrete, consider an alternative model in which the agent is known to share the principal's beliefs P. The principal proposes a contract w and recommends an effort e_i. The agent correctly assesses his payoff for e_i but uses the affine approximation

$$\begin{aligned} \frac{\widehat{\langle w, P_{i+1}\rangle} - w(\overline{y})}{\langle w, P_i\rangle - w(\overline{y})} &\equiv \frac{\overline{y} - \mu_{i+1}}{\overline{y} - \mu_i} && i < k, \\ \frac{\widehat{\langle w, P_{i-1}\rangle} - w(\underline{y})}{\langle w, P_i\rangle - w(\underline{y})} &\equiv \frac{\mu_{i-1} - \underline{y}}{\mu_i - \underline{y}} && i > k \end{aligned}$$

to evaluate his payoff under a marginal reduction in effort. He accepts the principal's recommendation if and only if the approximated payoff differential resulting from that effort reduction

$$\begin{aligned} \langle w, P_i\rangle - \widehat{\langle w, P_{i+1}\rangle} &= \frac{\mu_{i+1} - \mu_i}{\overline{y} - \mu_i}(\langle w, P_i\rangle - w(\overline{y})) && i < k, \\ \langle w, P_i\rangle - \widehat{\langle w, P_{i-1}\rangle} &= \frac{\mu_i - \mu_{i-1}}{\mu_i - \underline{y}}(\langle w, P_i\rangle - w(\underline{y})) && i > k \end{aligned} \tag{10}$$

matches or exceeds the effort cost differential $e_i - e_{i-1}$ or $e_i - e_{i+1}$, as appropriate. Unsurprisingly, the principal might as well pay the agent with an affine contract. To see why, let w be a contract of arbitrary form and suppose the principal's recommended effort e_i is incentive compatible. The principal's payoff for effort e_i and the agent's



(a) improved contract for e_2 (b) improved contract for e_3

Figure 5: Constructing an improving linear contract for belief type S

marginal payoff for effort e_i are each the same under the affine contract

$$\widehat{w}(z) \equiv \frac{\langle w, P_i\rangle - w(\overline{y})}{\overline{y} - \langle y, P_i\rangle}(\overline{y} - z) + w(\overline{y}) \qquad i < k,$$
$$\widehat{w}(z) \equiv \frac{\langle w, P_i\rangle - w(\underline{y})}{\langle y, P_i\rangle - \underline{y}}(z - \underline{y}) + w(\underline{y}) \qquad i > k$$

as they are under w.[11] Payoffs for both parties are as if the principal paid the agent with the affine contract $\widehat{w}$ rather than the original contract w.

The construction of the adversarial belief S is motivated by the approximating agent's incentive compatibility criterion, and for every contract w the actual marginal payoff

$$\langle w, S_i - S_{i+1}\rangle = \tau_i(\langle w, P_i\rangle - w(\overline{y})) = \frac{\mu_{i+1} - \mu_i}{\overline{y} - \mu_i}(\langle w, P_i\rangle - w(\overline{y})) \qquad i < k,$$
$$\langle w, S_i - S_{i-1}\rangle = \tau_i(\langle w, P_i\rangle - w(\underline{y})) = \frac{\mu_i - \mu_{i-1}}{\mu_i - \underline{y}}(\langle w, P_i\rangle - w(\overline{y})) \qquad i > k \qquad (11)$$

is identical to the linearized marginal payoff (10), as we illustrate in Figure 5.[12] The

[11]The coefficient on the first term of the improved contract $\widehat{w}$ is strictly positive because the original contract w satisfies the counterfactual agent's incentive compatibility criterion for effort e_i and the constant in the second term is nonnegative because w is a valid contract. Accordingly, $\widehat{w}$ satisfies limited liability and is therefore itself a well defined contract.

[12]The diagrams in Figure 5 are generated to scale by the data $\mathcal{Y} \equiv [-1, 1]$, $w \equiv y^3 - y/2$, $n \equiv 3$,

subtlety inherent in recovering the approximating agent's behavior within the expected utility maximization framework is the need for that equivalence to hold simultaneously across distinct effort recommendations. In particular, while the approximating agent's payoffs are identical to his expected utility under beliefs A with

$$A(e_i) \equiv P(e_i), \qquad A(e_{i+1}) \equiv \frac{\mu_{i+1} - \mu_i}{\overline{y} - \mu_i}\delta(\overline{y}) + \frac{\overline{y} - \mu_{i+1}}{\overline{y} - \mu_i}P(e_i) \qquad i \leq k,$$
$$A(e_i) \equiv P(e_i), \qquad A(e_{i-1}) \equiv \frac{\mu_i - \mu_{i-1}}{\mu_i - \underline{y}}\delta(\underline{y}) + \frac{\mu_{i-1} - \underline{y}}{\mu_i - \underline{y}}P(e_i) \qquad i \geq k,$$

the competing specifications

(i) $A(e_i) \equiv P(e_i)$ if the principal recommends effort e_i,

(ii) $A(e_i) \equiv (1-\lambda)\delta(\underline{y}) + \lambda P(e_{i+1})$ if she recommends effort e_{i+1} for $i \geq k$,

(iii) $A(e_i) \equiv (1-\lambda)\delta(\overline{y}) + \lambda P(e_{i-1})$ if she recommends effort e_{i-1} for $i \leq k$

are generically contradictory in cases of our model with more than two actions. Accordingly, while the linearly approximated marginal payoffs (10) coincide with the actual marginal payoffs (11), the absolute payoffs typically differ, as in Figure 5.

4 Beyond common knowledge of the relationship between effort and expected output

The assumption that the relationship between effort and expected output is common knowledge is restrictive in two ways. First, the principal and the agent agree about that relationship. Second, that agreement is understood by the principal. In this section, we show that our strongest result (Theorem 3) persists under relaxations of both.

Assumption 1*. *(Supermodularity) There exists a complete lattice $(\Theta, \leq)$ and a supermodular function $f : \mathcal{E} \times \Theta \to \mathcal{Y}$ such that $\mathscr{A} = \{f(\cdot|\theta)|\theta \in \Theta\}$.*

$(\mu_1, \mu_2, \mu_3) \equiv (-\sqrt{6}/6, 0, \sqrt{6}/6)$, $\langle w, P_2\rangle \equiv -3/16$, and $\langle w, P_3\rangle \equiv 1/6$. The low-effort wage $\langle w, P_1\rangle$ is set arbitrarily with $(\mu_1, \langle w, P_1\rangle) \in \Phi(w)$ and the payoffs $\langle w, S_1\rangle, \langle w, S_2\rangle, \langle w, S_3\rangle$ are computed according to the definition of S.

Assumption 1* provides for the existence of a plausible belief about expected output that is minimal in the sense of increasing differences. We write a^* for the element of $\mathscr{A}$ that satisfies

$$a(\max\{e, e'\}) - a(\min\{e, e'\}) \geq a^*(\max\{e, e'\}) - a^*(\min\{e, e'\}) \tag{12}$$

for all plausible beliefs about expected output $a \in \mathscr{A}$. This belief plays a central role in our analysis, and we make two comments about the assumption. First, if $\mathscr{A}$ contains only a single element, then Assumption 1* is vacuously satisfied. Second, our arguments go through without modification under the weaker assumption that $\mathscr{A}$ has a minimal element in the sense of increasing differences, as in (12).

Definition 1*. Effort e is *efficient** if $e \leq \varepsilon$ for every point ε such that $(\varepsilon, a^*(e))$ belongs to the convex hull of graph(a^*).

As before, we enumerate the efficient* efforts $e_1^*, ..., e_n^*$ so that $a_i^* \equiv a^*(e_i^*)$ is strictly increasing in i. We label as l the index associated with the zero effort action and use the shorthand $P_i^* \equiv P(e_i^*)$ for the principal's beliefs, $p_i^* \equiv p(e_i^*)$ for the expected value of the principal's beliefs, and $\pi_i^* \equiv \pi(e_i^*)$ for the output component of the principal's payoff, noting the dependency of the index $i = 1, ..., n$ on the identity of a^*. With that notation in hand, we proceed to our construction. Define weights τ_i^*, remainder r^* by

$$\tau_i^* \equiv \begin{cases} \dfrac{a_{i+1}^* - a_i^*}{\overline{y} - p_i^*} & i < l, \\ 0 & i = l, \\ \dfrac{a_i^* - a_{i-1}^*}{p_i^* - \underline{y}} & i > l; \end{cases} \qquad r^* \equiv \frac{a_l^* - \underline{y}\sum_{i>l} \tau_i^* - \overline{y}\sum_{i<l} \tau_i^*}{1 - \sum_i \tau_i^*}$$

and beliefs S_i^* by

$$S^*(e_i) \equiv S_i^* \equiv \begin{cases} \sum_{j<i} \tau_j^* \delta(\overline{y}) + \sum_{j=i}^{l} \tau_j^* P_j^* + \sum_{j>l} \tau_j^* \delta(\underline{y}) + (1 - \sum_j \tau_j^*)\delta(r^*) & i \leq l, \\ \sum_{j<l} \tau_j^* \delta(\overline{y}) + \sum_{j=l}^{i} \tau_j^* P_j^* + \sum_{j>i} \tau_j^* \delta(\underline{y}) + (1 - \sum_j \tau_j^*)\delta(r^*) & i \geq l. \end{cases}$$

Extend the definition to the inefficient* efforts e by choosing $\xi^e \in \Delta(\{1, ..., n\})$

to satisfy $\sum_i \xi_i^e a_i^* = a^*(e), \sum_i \xi_i^e e_i^* < e$ and assigning $S^*(e) \equiv \sum_i \xi_i^e S_i^*$. Toward the statement of conditions under which S^* is plausible, define

$$\kappa^* \equiv \max\left(\max_{i<l} \frac{\overline{y} - a_i^*}{\overline{y} - p_i^*}, \max_{i>l} \frac{a_i^* - \underline{y}}{p_i^* - \underline{y}}, \right), \qquad \varphi^* \equiv \frac{\sqrt{4(\kappa^*)^2 + 1} + 2\kappa^* - 1}{2}$$

and note that φ^* reduces to $(1/2)(1+\sqrt{5})$ in the $a^* = p$ case, as in Assumption 2.

Assumption 2*. *(Broad* support) First, every belief A with $\langle y, A(e)\rangle = a^*(e)$ for every effort e is plausible. Second, the set of outputs $\mathcal{Y} = [\underline{y}, \overline{y}]$ satisfies*

$$\frac{a_1^* - \underline{y}}{a_n^* - a_1^*} \geq \varphi^*, \qquad \frac{\overline{y} - a_n^*}{a_n^* - a_1^*} \geq \varphi^*.$$

Together, Assumptions 1*, and 2* yield a proper generalization of our earlier model with common knowledge of the relationship between effort and expected output. In order to state our main results for this section, it will again be useful to introduce the linear contracts of interest. Define for each i

$$w_i^*(y) \equiv \begin{cases} \frac{e_i^* - e_{i+1}^*}{a_{i+1}^* - a_i^*}(\overline{y} - y) & i < l, \\ 0 & i = l, \\ \frac{e_i^* - e_{i-1}^*}{a_i^* - a_{i-1}^*}(y - \underline{y}) & i > l; \end{cases} \quad t_i^* \equiv \begin{cases} \frac{e_i^* - e_{i+1}^*}{a_{i+1}^* - a_i^*}(\overline{y} - p_i^*) & i < l, \\ 0 & i = l, \\ \frac{e_i^* - e_{i-1}^*}{a_i^* - a_{i-1}^*}(p_i^* - \underline{y}) & i > l; \end{cases} \quad \Pi_i^* \equiv \pi_i^* - t_i^*,$$

let c_i^* be any incentive compatible tiebreaking rule that breaks ties in favor of e_i^* wherever possible, and set $W_i^* \equiv (w_i^*, c_i^*)$. We first consider specifications of our model in which a^* is the lone element of $\mathscr{A}$. Under no additional assumptions, a proper generalization of Theorem 3 obtains.

Theorem 4. *Let Assumptions 1*, 2* hold and suppose that the agent's beliefs about expected output are known to be a^*. First, the principal's problem has a saddle point. Second, (W_i^*, S^*) is a saddle point if and only if $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$. Second, if the principal's beliefs have full support, then W solves the principal's problem only if $W = W_i^*$ for some index i with $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$.*

We establish the result in steps. First, we show that S^* is well defined and plausible. Second, we show that the principal's payoff in the counterfactual in which she knows

the agent's beliefs are S^* is bounded above by $\max(\Pi_1^*, ..., \Pi_n^*)$. Third, we verify that W_i^* guarantees the principal payoff Π_i^* across the set of plausible beliefs. Fourth, and finally, we show that nonlinear contracts are suboptimal. Both the structure of the overall argument and the details of the individual steps closely mirror the logic of Theorem 3.

Setting aside further technical elaboration until the Appendix, the premise that the principal knows the agent's beliefs about expected output is itself informationally demanding, if not especially so when her own beliefs differ from his. Under mild regularity conditions, our result extends to a much richer class of environments.

Assumption 3. *(Regularity) First, the principal's beliefs about expected output p and the agent's beliefs about expected output a are nondecreasing in effort for every plausible belief about expected output a. Second, the principal's utility function satisfies $\pi = p$.*

Monotonicity of the agent's beliefs works together with our supermodularity assumption to ensure all agent types exert no less effort than type S^* when paid with a linear contract. At the same time, monotonicity of the principal's beliefs and full linearity of her utility function ensure her payoff is nondecreasing in effort.

Theorem 5. *Let Assumptions 1*, 2*, 3 hold. First, the principal's problem has a saddle point. Second, (W_i^*, S^*) is a saddle point if and only if $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$. Second, if the principal's beliefs have full support, then W solves the principal's problem only if $W = W_i^*$ for some index i with $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$.*

Altogether, Theorem 5 provides a sharp optimality result for specifications of our model in which there is (i) monotonicity of beliefs about expected output in effort; (ii) linearity of the principal's utility function; and (iii) a minimal belief in the sense of increasing differences. All three assumptions are standard and, in our view, at least the first and the third are broadly permissive.

5 Related literature

Robust contracts and mechanisms This paper complements the literature on robust contracting with uncertain production technologies initiated by Chassang

(2013) and Carroll (2015) and elaborated on by Dai and Toikka (2022); Walton and Carroll (2022); Kambhampati (2023, 2024); Burkett and Rosenthal (2024a); Antic and Georgiadis (2024); Marku, Ocampo, and Tondji (2024); Liu (2024); Kambhampati, Peng, Tang, Toikka, and Vohra (2025). While the defining premise of that literature is the principal's incomplete understanding of the agent's capabilities, our principal knows both the set of actions available to the agent and also the distribution of output produced by each of those actions. This knowledge protects her from the all-or-nothing risk taking that renders convex contracts suboptimal in the baseline unknown technology model, and our optimal linear bonus contracts are only sometimes fully linear.

The methods we use in proving Theorem 1 and Theorem 2 are inspired by the elegant applications of convex analysis in the aforementioned Carroll (2015). There, the author uses the separating hyperplane theorem to improve given contracts of arbitrary form into linear contracts that yield better guarantees. For our part, we use the supporting hyperplane theorem to construct improving linear contracts in Theorem 1 and improving linear bonus contracts in Theorem 2. The high level distinction between the two sets of arguments is that while Carroll's approach is based around improving the agent's payoff for desirable actions while maintaining his payoff for undesirable actions, our approach is based around tightening the range of payoffs associated to the agent's actions while simultaneously preserving or degrading his wages. While improving the agent's payoff for "good" actions is valuable to Carroll's principal because doing so endogenously improves the productivity of the "bad" actions that drive her guarantees, our agent's actions are fixed and transferring additional utility to him only serves to increase our principal's costs.

Setting aside our first two major results, our saddle point theorems have no counterpart in Carroll (2015) or elsewhere in the ensuing literature because the uncertain technology problem in which the principal is constrained to the deterministic use of a single contract does not have a saddle point except in a narrow class of special cases, as Carroll shows.[13] While Carroll does prove directly that screening does

[13]Kambhampati, Peng, Tang, Toikka, and Vohra (2025) identify a saddle point in the expanded formulation of Carroll's problem that allows for randomization over contracts. Conversely, as Carroll shows, the formulation of the problem in which the principal is constrained to deterministic contracts does not have a saddle point. By contrast, we identify saddle points of our problem in which

not improve that principal's guarantee, other authors have since established that randomization (Kambhampati (2023); Kambhampati, Peng, Tang, Toikka, and Vohra (2025)) and experimentation (Liu (2024)) do. In contrast, we identify broad conditions under which none of the aforementioned procedures yield a better payoff guarantee than the best linear contract. The significance of this feature of our model is admittedly subjective, and there are practical reasons why a real-world principal in a Carroll-like environment might wish to use a single contract even if complex mechanisms do yield better guarantees.

Separately, our richness condition is related to the richness condition that appears in Walton and Carroll (2022). There, the authors suppose that the set of output distributions $\Phi(w)$ taken in best response to contract w has the property that if (i) F is a potential best response; (ii) expected output under F' and F is the same; and (iii) expected transfers under F' are at least expected transfers under F, then F' is also a potential best response. Aside from minor technicalities pertaining to (ii), we differ in two ways. First, we work with agent beliefs A rather than shared beliefs F. Second, our uncertainty set $\mathcal{A}$ is primitive while theirs $\Phi(w)$ is derived from primitives. Hence, there is no analogue to (iii) in our setting.

Elsewhere in the robust contracts literature, a pair of related papers (Lopomo, Rigotti, and Shannon (2011); Burkett and Rosenthal (2024b)) study a model in which the agent is uncertain about his own production technology and demands to be paid with a contract under which the recommended level of effort is unambiguously incentive compatible. As in our paper, the principal knows the agent's production technology; unlike in our paper, the agent does not. In contrast with our own results, both studies find that optimal contracts are coarse.

Finally, while this paper is to the best of our knowledge the first to study belief-robust incentive contracting, several earlier studies consider robustness to beliefs in mechanism design problems with multiple agents (Bergemann and Morris (2005); Chung and Ely (2007); Brooks and Du (2021)).

the constituent optimal contract does not involve randomization. This distinction is significant. In our model, deterministic contracts are optimal in expanded design spaces; in Carroll's model, deterministic contracts are only optimal in expanded design spaces for a proper subset of the family of cases in which the 0 contract is optimal. Kambhampati, Peng, Tang, Toikka, and Vohra do clarify that randomized screening menus do not yield better guarantees than the best random contract.

Contracting with heterogeneous beliefs One approach to relaxing the common prior assumption that is different than our own is to instead assume the principal and the agent have different beliefs but the agent's beliefs are known to the principal. This premise has been explored both in the context of moral hazard (Karni (2008)) and separately in the extensive behavioral contract theory literature (see Kőszegi (2014) for an overview), where it has been shown for instance that calculating firms are in a position to benefit from overconfidence in an employee (de la Rosa (2011)) or naïveté in a borrower (Heidhues and Kőszegi (2010)). While much of that literature interprets the principal's beliefs as "correct" and the agent's beliefs as "biased", we have deliberately avoided assigning primacy to either parties' beliefs and we have eschewed welfare calculations altogether. Instead, our emphasis lies on the principal's uncertainty about the agent's beliefs, rather than on disagreement per se.

6 Conclusions

This paper studies the design of robust incentives for an agent with unknown beliefs about the output generating process. We show that linear bonus contracts are optimal in general and that linear contracts are optimal in special cases. Further, we show that if the range of outputs is large relative to expected productivity, then the principal's problem has a saddle point and optimal contracts are again linear. In doing so, we provide foundations for paying the agent with simple contracts on the basis of robustness to uncertainty about beliefs.

Our analysis here suggests a few avenues for future work. First, our optimality results leave open the possibility that our optimal contracts are dominated by more complicated mechanisms that employ screening or randomization. It remains unclear whether or not the characterization of undominated incentive schemes is a tractable problem, and similarly whether or not its solutions are of interest. Second, we have neither confirmed nor refuted the existence of saddle points in specifications of the principal's problem that fail Assumption 2, 2*. To the extent that screening or randomization might be impractical in certain real-world applications, a positive result for those specifications seems desirable. Third, and finally, our assumption that the agent's preferences are quasilinear in effort might be viewed as restrictive.

While the core improvement arguments in Lemmas 3, 5 that respectively support Theorems 1, 2 generalize readily to any environment in which the agent's preferences are monotonically increasing in expected transfers, our notions of efficiency and our accompanying characterization of optimal linear contracts do not extend to those alternative environments. Furthermore, because the improvements that support Theorems 3–5 are based around preserving the agent's marginal payoff for the intended effort but not necessarily his absolute payoff, it is unclear if those results obtain under more flexible restrictions to the agent's preferences.

References

Antic, N. and G. Georgiadis (2024, 08). Robust Contracts: A Revealed Preference Approach. *The Review of Economics and Statistics*, 1–31.

Bergemann, D. and S. Morris (2005). Robust mechanism design. *Econometrica 73*(6), 1771–1813.

Brooks, B. and S. Du (2021). Optimal auction design with common values: An informationally robust approach. *Econometrica 89*(3), 1313–1360.

Burkett, J. and M. Rosenthal (2024a). Data-driven contract design. *Journal of Economic Theory 221*, 105900.

Burkett, J. and M. Rosenthal (2024b). Statistical uncertainty and coarse contracts. *Journal of Economic Theory 220*, 105876.

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review 105*(2), 536–563.

Chassang, S. (2013). Calibrated incentive contracts. *Econometrica 81*(5), 1935–1971.

Chung, K. S. and J. C. Ely (2007). Foundations of dominant-strategy mechanisms. *Review of Economic Studies 74*(2), 447–476.

Dai, T. and J. Toikka (2022). Robust incentives for teams. *Econometrica 90*(4), 1583–1613.

de la Rosa, L. E. (2011). Overconfidence and moral hazard. *Games and Economic Behavior 73*(2), 429–451.

Eggleston, H. G. (1958). *Convexity*. Cambridge University Press.

Heidhues, P. and B. Kőszegi (2010). Exploiting naivete about self-control in the credit market. *American Economic Review 100*(5), 2279–2303.

Innes, R. D. (1990). Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory 52*(1), 45–67.

Ito, K. (2014). Do consumers respond to marginal or average price? evidence from nonlinear electricity pricing. *American Economic Review 104*(2), 537–563.

Kambhampati, A. (2023). Randomization is optimal in the robust principal-agent problem. *Journal of Economic Theory 207*, 105585.

Kambhampati, A. (2024). Robust performance evaluation of independent agents. *Theoretical Economics 19*(3), 1151–1184.

Kambhampati, A., B. Peng, Z. G. Tang, J. Toikka, and R. Vohra (2025). Randomization and the robustness of linear contracts. Unpublished paper.

Karni, E. (2008). On optimal insurance in the presence of moral hazard. *The Geneva Risk and Insurance Review 33*(1), 1–18.

Kőszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Kreps, D. M. (2012). *Microeconomic foundations I: Choice and competitive markets*. Princeton University Press.

Liu, C. (2024). Robust contracts with exploration. Unpublished paper.

Lopomo, G., L. Rigotti, and C. Shannon (2011). Knightian uncertainty and moral hazard. *Journal of Economic Theory 146*(3), 1148–1172.

Marku, K., S. Ocampo, and J.-B. Tondji (2024). Robust contracts in common agency. *The RAND Journal of Economics 55*(2), 199–229.

Martimort, D. and L. A. Stole (2020). Nonlinear pricing with average-price bias. *American Economic Review: Insights 2*(3), 375–396.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy 11*, 227–253.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons.

Walton, D. and G. Carroll (2022). A general framework for robust contracting models. *Econometrica 90*(5), 2129–2159.

A Section 2

Proof of Lemma 1. Let w be an arbitrary contract. We prove both directions directly. First, let $a \in \mathscr{A}$ and let payoff vector $v : \mathcal{E} \to \mathbb{R}_+$ satisfy $(a(e), v(e)) \in \Phi(w)$ for every effort e. Because w is continuous and $\mathcal{Y}$ is an interval and therefore connected, $\Phi(w)$ is itself a connected subset of $\mathbb{R}^2$. In turn, the Fenchel-Bunt strengthening of Caratheodory's theorem for connected sets (see Eggleston (1958) Theorem 18(ii)) yields vectors $\lambda : \mathcal{E} \to [0,1]$ and $a_0, a_1 : \mathcal{E} \to \mathcal{Y}$ such that $a(e) = (1-\lambda(e))a_0(e) + \lambda a_1(e)$ and $v(e) = (1-\lambda(e))w(a_0(e)) + \lambda(e)w(a_1(e))$. The belief A with $A(e) \equiv (1-\lambda(e))\delta(a_0(e)) + \lambda(e)\delta(a_1(e))$ is plausible and satisfies $\langle w, A(e)\rangle = v(e)$ for all e.

Second, let A be any plausible belief. Let a be the output vector $e \mapsto \langle y, A(e)\rangle$ and v the wage valuation $e \mapsto \langle w, A(e)\rangle$. If $A(e)$ has finite support, then it follows immediately from the definition of $\Phi(w)$ that $(a(e), v(e)) \in \Phi(w)$. If $A(e)$ does not have finite support, recall that the set of finitely supported distributions is dense in $\Delta(\mathcal{Y})$ and the payoff set $\Phi(w)$ is closed. The infinite case follows from the finite cases via limits. □

Proof of Fact 1. First, if w is affine then $(y,t) \in \Phi(w)$ if and only if $t = w(y)$. Second, if w is not affine then for every interior y there exist outputs $y_0, y_1 \in \mathcal{Y}$ and constant $\lambda \in (0,1)$ such that $y = (1-\lambda)y_0 + \lambda y_1$ and $w(y) \neq (1-\lambda)w(y_0) + \lambda w(y_1)$. Accordingly, $\Phi(w)$ contains distinct pairs $(y, w(y))$ and $(y, (1-\lambda)w(y_0) + \lambda w(y_1))$. □

Proof of Fact 2. If w is affine, then $\Phi(w)$ has an empty interior per Fact 1. Conversely, if w is nonaffine then write $t^0 \equiv \min\{t|(\langle y, F\rangle, t) \in \Phi(w)\}$ and $t^1 \equiv \max\{t|(\langle y, F\rangle, t) \in \Phi(w)\}$. The supporting hyperplane theorem provides a support line ℓ^0 to $\Phi(w)$ at $(\langle y, F\rangle, t^0)$ and a support line ℓ^1 to $\Phi(w)$ at $(\langle y, F\rangle, t^1)$. Let $T^0 \subset \mathcal{Y}$ be the projection of $\text{graph}(w) \cap \ell^0$ onto its first coordinate and $T^1 \subset \mathcal{Y}$ the projection of $\text{graph}(w) \cap \ell^1$ onto its first coordinate. First, per our choice of ℓ^0, ℓ^1 we have (i) $\langle w, F\rangle = t^0 \iff \text{supp}(F) \subset T^0$ and (ii) $\langle w, F\rangle = t^1 \iff \text{supp}(F) \subset T^1$. Second, because w is continuous and nonaffine there exists an open set $O^0 \subset \mathcal{Y}$ such that $T^0 \cap O^0 = \emptyset$ and an open set $O^1 \subset \mathcal{Y}$ such that $T^1 \cap O^1 = \emptyset$. Accordingly, because F has full support, we have $t^0 < \langle w, F\rangle < t^1$ as claimed. □

B Section 3

B.1 Section 3.1

Proof of Lemma 2. Suppose e is efficient and write $m \equiv \mu(e)$. Write $S \subset \mathbb{R}^2$ for the convex hull of the graph of μ and $T \subset \mathbb{R}^2$ for the set of all pairs (ε, m) with $\varepsilon < e$. First, S and T are disjoint per our choice of e. Second, S, T are both apparently convex. Third, (e, m) belongs to both the boundary of both S and T. In turn, the separating hyperplane theorem yields a normal vector $(\eta_e, \eta_m) \neq (0, 0)$ such that $\eta_e e_S + \eta_m m_S \geq \eta_e e + \eta_m m \geq \eta_e e_T + \eta_m m_T$ for all $(e_S, m_S) \in S$ and for all $(e_T, m_T) \in T$. Per our choice of T, we have $\eta_e \geq 0$. There are two cases to consider. First, if $\eta_e = 0$ then $\eta_m \neq 0$ and thus e either maximizes or minimizes the expected output map μ. In that case, because e is efficient, e is uniformly incentive compatible under contracts of the form $w(y) = \alpha y + \beta$ with α appropriately signed and sufficiently large in magnitude. Second, if $\eta_e > 0$ then e is uniformly incentive compatible under contracts of the form $w(y) = (-\eta_m/\eta_e)y + \beta$.

Conversely, suppose e is inefficient. By definition, there exists a convex combination $\gamma \in \Delta(\mathcal{E})$ such that $\sum_{\mathcal{E}} \gamma(\varepsilon)\mu(\varepsilon) = \mu(e)$ and $\sum_{\mathcal{E}} \gamma(\varepsilon)\varepsilon < e$. Consider the belief

$$A(\varepsilon) \equiv \frac{\overline{y} - \mu(\varepsilon)}{\overline{y} - \underline{y}}\delta(\underline{y}) + \frac{\mu(\varepsilon) - \underline{y}}{\overline{y} - \underline{y}}\delta(\overline{y})$$

and note that (i) A is plausible, (ii) $\langle w, A(e)\rangle - e = \sum_{\varepsilon\in\mathcal{E}} \gamma(\varepsilon)\langle w, A(\varepsilon)\rangle - e < \sum_{\varepsilon\in\mathcal{E}} \gamma(\varepsilon)(\langle w, A(\varepsilon)\rangle - \varepsilon)$ for every contract w. Apparently, there are no contracts under which e is incentive compatible for A. □

Proof of Lemma 4. There are two cases to consider. If $i = k$ then w_i is the zero contract and the claim follows immediately from limited liability. Alternatively, suppose $i \neq k$ and consider linear contract $\alpha y + \beta$. Per Lemma 2, it is sufficient to verify incentive compatibility against other efficient efforts. Accordingly, for all j distinct from i write $\alpha_j \equiv (\mu_i - \mu_j)^{-1}(e_i - e_j)$ and note that the downward conditions $\alpha \geq \alpha_j$ for all $j < i$ and the upward conditions $\alpha \leq \alpha_j$ for all $j > i$ are jointly necessary and sufficient for incentive compatibility. Further, because the restriction of μ to the set of efficient efforts is by definition the restriction of some concave function on the convex hull of the complete set of efforts $\mathcal{E}$, α_j is nondecreasing in j. Consequently, the adjacent conditions $\alpha_{i+1} \geq \alpha \geq \alpha_{i-1}$ are both necessary and sufficient. If $i > k$, the best contract in that set is $\alpha_{i-1}(y - \underline{y}) = w_i$. If $i < k$, the best contract in that set is $-\alpha_{i+1}(\overline{y} - y) = w_i$. □

Proof of Proposition 1. We claim that if W does not uniformly incentivize one of the two efforts then $\Pi(W) < \Pi_k$. Once that has been established, the result is a corollary of Theorem 1. Proceeding, suppose $W = (w, c)$ is such that there exist plausible beliefs A, B with $c(A) = \min \mathcal{E}$ and $c(B) = \max \mathcal{E}$. First, because $\min \mathcal{E}$ is strictly incentive compatible under contract w_k for every plausible belief and $c(B) \neq \min \mathcal{E}$, we have $w \neq w_k$. Second, because the principal's beliefs have full support and w is continuous and satisfies limited liability, we have $\langle w, P(0)\rangle > \langle w_k, P(0)\rangle$. Consequently, $\Pi(W) \leq \Pi(W|A) = \pi_k - \langle w, P(0)\rangle < \pi_k - \langle w_k, P(0)\rangle = \Pi_k$. □

B.2 Section 3.2

Proof of Theorem 2. Write Π^∞ for the supremum of the objective in the principal's problem, noting that Π^∞ is finite because the set of efforts $\mathcal{E}$ is finite. We claim that there exists a contract w and an incentive compatible recommendation c such that $W = (w, c)$ satisfies $\Pi(W) = \Pi^\infty$. The result then follows from Lemma 5. If $\Pi^\infty = \Pi_k$, then W_k is a solution to the principal's problem per Theorem 1. Accordingly, we restrict attention for the rest of the proof to the more interesting cases in which

$\Pi^\infty > \Pi_k$. Parameterize the set of linear bonus contracts by $w^{\alpha,x} \equiv \max(\alpha(y-x),0)$ with $\alpha \in \mathbb{R}, x \in \mathcal{Y}$ and let $c^{\alpha,x}$ break ties in favor of the principal. Define $E(\alpha,x) \equiv \cup_{A\in\mathcal{A}} c^{\alpha,x}(A)$ and $\Pi(\alpha,x) \equiv \Pi((w^{\alpha,x}, c^{\alpha,x}))$ for the corresponding set of recommended efforts and principal payoffs, respectively.

Lemma 5 provides for the existence of a sequence of parameters (α_j, x_j) with $\Pi(\alpha_j, x_j) \to \Pi^\infty$. For the purposes of readability, we write $w^j \equiv w^{\alpha_j,x_j}, c^j \equiv c^{\alpha_j,x_j}$ and implicitly pass to subsequences wherever necessary. First, because the set of efforts $\mathcal{E}$ is finite, there exists a fixed set of efforts E^∞ such that $E(\alpha_j, x_j) = E^\infty$. Second, because $\mathcal{Y}$ is compact, $x_j \to x^\infty$ for some output x^∞. Third, if $\alpha_j \to \alpha^\infty$ for some $\alpha^\infty \in \mathbb{R}$, then $\Pi(\alpha^\infty, x^\infty) = \Pi^\infty$ because the principal's payoff for each effort is continuous in (α, x).

We show by contradiction that α_j must indeed converge. Toward that end, suppose to the contrary that α_j is unbounded and consider the plausible belief A defined by

$$A(e) \equiv \begin{cases} \delta(\mu(e)) & e > 0, \\ \frac{\overline{y}-\mu(0)}{\overline{y}-\underline{y}}\delta(\underline{y}) + \frac{\mu(0)-\underline{y}}{\overline{y}-\underline{y}}\delta(\overline{y}) & e = 0. \end{cases}$$

There are two cases to consider. First, if $\underline{y} < x^\infty < \overline{y}$ then $\langle w^j, A(0)\rangle \to +\infty$. In turn, incentive compatibility implies $\langle w^j, A(c^j(A))\rangle \to +\infty$. Because Jensen's inequality implies $\langle w^j, P(e)\rangle \geq \langle w^j, A(e)\rangle$ for all positive efforts e, we have either $\langle w^j, P(c^j(A))\rangle \to +\infty$ or $0 \in E^\infty$. Both subcases contradict $\Pi^\infty > \Pi_k$. Second, if $x^\infty = \underline{y}$ or if $x = \overline{y}$, then either (i) $\langle w^j, A(e)\rangle \to +\infty$ for all nonzero efforts or (ii) $\langle w^j, A(e)\rangle \to 0$ for all nonzero efforts. In subcase (i), the situation is the same as in the $\underline{y} < x^\infty < \overline{y}$ case. In subcase (ii), $0 \in E^\infty$. Both subcases again contradict $\Pi^\infty > \Pi_k$, and so we conclude α_j indeed converges to a finite limit α^∞. □

B.3 Section 3.3

Proof of Lemma 6. We show first that S is composed of valid probability distributions and second that S is consistent with common knowledge of the relationship between effort and expected output.

Part 1 First, from the definition of τ_i and the approximation $\ln(x) \leq x - 1$ we have

$$\sum_{i<k} \tau_i < \int_{\mu_1}^{\mu_k} \frac{1}{\overline{y} - x}\, \mathrm{d}x = \ln\left(\frac{\overline{y} - \mu_1}{\overline{y} - \mu_k}\right) \leq \frac{\mu_k - \mu_1}{\overline{y} - \mu_k},$$
$$\sum_{i>k} \tau_i < \int_{\mu_k}^{\mu_m} \frac{1}{x - \underline{y}}\, \mathrm{d}x = \ln\left(\frac{\mu_m - \underline{y}}{\mu_k - \underline{y}}\right) \leq \frac{\mu_m - \mu_k}{\mu_k - \underline{y}}.$$

Second, from $\underline{y} < \mu_1 \leq \mu_k \leq \mu_m < \overline{y}$ we have

$$\frac{\mu_k - \mu_1}{\overline{y} - \mu_k} \leq \frac{\mu_k - \mu_1}{\overline{y} - \mu_m} = \frac{\mu_k - \mu_1}{\mu_m - \mu_1}\frac{\mu_m - \mu_1}{\overline{y} - \mu_m} \leq \frac{\mu_k - \underline{y}}{\mu_m - \underline{y}}\frac{\mu_m - \mu_1}{\overline{y} - \mu_m},$$
$$\frac{\mu_m - \mu_k}{\mu_k - \underline{y}} \leq \frac{\mu_m - \mu_k}{\mu_1 - \underline{y}} = \frac{\mu_m - \mu_k}{\mu_m - \mu_1}\frac{\mu_m - \mu_1}{\mu_1 - \underline{y}} \leq \frac{\overline{y} - \mu_k}{\overline{y} - \mu_1}\frac{\mu_m - \mu_1}{\mu_1 - \underline{y}}.$$

Third, define $\Delta \equiv \mu_m - \mu_1, \varphi \equiv (1/2)(1 + \sqrt{5}), \underline{\varepsilon} \equiv (\mu_1 - \underline{y}) - \varphi\Delta, \overline{\varepsilon} \equiv (\overline{y} - \mu_m) - \varphi\Delta$ and note Assumption 2 implies $\underline{\varepsilon}, \overline{\varepsilon} \geq 0$. From $\varphi \approx 1.6 > 1$ and $(1 + \varphi)\varphi = (1 + 2\varphi)$,

$$\frac{\mu_k - \underline{y}}{\mu_m - \underline{y}}\frac{\mu_m - \mu_1}{\overline{y} - \mu_m} = \frac{(\mu_k - \underline{y})\Delta}{(\Delta + \varphi\Delta + \underline{\varepsilon})(\varphi\Delta + \overline{\varepsilon})} \leq \frac{\mu_k - \underline{y}}{\Delta(1 + \varphi)\varphi + \underline{\varepsilon} + \overline{\varepsilon}} = \frac{\mu_k - \underline{y}}{\overline{y} - \underline{y}},$$
$$\frac{\overline{y} - \mu_k}{\overline{y} - \mu_1}\frac{\mu_m - \mu_1}{\mu_1 - \underline{y}} = \frac{(\overline{y} - \mu_k)\Delta}{(\varphi\Delta + \underline{\varepsilon})(\Delta + \varphi\Delta + \overline{\varepsilon})} \leq \frac{\overline{y} - \mu_k}{\Delta(1 + \varphi)\varphi + \underline{\varepsilon} + \overline{\varepsilon}} = \frac{\overline{y} - \mu_k}{\overline{y} - \underline{y}}.$$

Altogether, we obtain

$$\sum_{i<k} \tau_i < \frac{\mu_k - \underline{y}}{\overline{y} - \underline{y}}, \tag{13}$$
$$\sum_{i>k} \tau_i < \frac{\overline{y} - \mu_k}{\overline{y} - \underline{y}}. \tag{14}$$

Two facts follow. First, (13) and (14) immediately imply $\sum_i \tau_i < 1$. Because $\tau_i \geq 0$ for all i by construction, the vector $(\tau_1, ..., \tau_m; 1 - \sum_i \tau_i)$ therefore specifies a valid convex combination. Second, (13) implies $r > \underline{y}$ and (14) implies $r < \overline{y}$. Consequently, the remainder distribution $\delta(r)$ is itself well defined. Altogether, we conclude $S(e)$ is a well defined element of $\Delta(\mathcal{Y})$ for each efficient effort e. This extends to the inefficient efforts e by construction.

Part 2 We claim that S_i has mean μ_i for every efficient effort e_i. This obtains immediately for $i = k$. Otherwise, from our choice of remainder r we have

$$\langle y, S_i \rangle = \begin{cases} \sum\limits_{i \leq j < k} \tau_j(\mu_j - \overline{y}) + \mu_k & i < k, \\ \sum\limits_{k < j \leq i} \tau_j(\mu_j - \underline{y}) + \mu_k & i > k \end{cases}$$

and from our choice of weights τ_i we have

$$\begin{aligned} \sum_{i \leq j < k} \tau_j(\mu_j - \overline{y}) + \mu_k &= \sum_{i \leq j < k} (\mu_j - \mu_{j+1}) + \mu_k = \mu_i, \\ \sum_{k < j \leq i} \tau_j(\mu_j - \underline{y}) + \mu_k &= \sum_{k < j \leq i} (\mu_j - \mu_{j-1}) + \mu_k = \mu_i. \end{aligned}$$

Accordingly, S_i has mean μ_i. In turn, because we chose γ^e to satisfy $\sum_i \gamma_i^e \mu_i = \mu(e)$ for each inefficient effort e, distribution $S(e)$ has mean $\mu(e)$ for each inefficient effort e and thus S is plausible. □

Proof of Lemma 8. Let $W = (w, c)$ be such that w is not a linear contract. First, if w does not have minimum payment 0 then $\tilde{W} \equiv (\tilde{w}, c)$ with $\tilde{w} \equiv w - \min_z w(z)$ satisfies $\Pi(\tilde{W}) > \Pi(W)$. Second, if $c(A) = 0$ for any plausible belief A then $\Pi(W_k) > \Pi(W)$, as in the proof of Proposition 1. Accordingly, suppose to the contrary that $c(A) \neq 0$ for all plausible A and that w has minimum payment 0. Write $E \equiv \arg\max_{e \in \mathcal{E}} \langle w, S(e) \rangle - e$ for the set of payoff-maximizing effort choices for belief S and note that every effort in E is efficient per Lemma 7.

Proceeding, because the principal's beliefs P_i have full support and S_i includes factor P_i for every $i \neq k$, S_i itself has full support for every $i \neq k$. In turn, Fact 2 implies that for every effort $e \in E$ there exists a plausible belief under which e is uniquely incentive compatible. This has two implications. First, it follows immediately that $\Pi(W) \leq \min_{e \in E} \pi(e) - \langle w, P(e) \rangle$. Second, because we have already restricted attention to cases with $c(A) > 0$ for all A, it must be that $0 \notin E$ and hence that E contains at least one nonzero effort. There are two cases to consider. Suppose $j > k$ for some j with $e_j \in E$ and set $i \equiv \min\{j | e_j \in E, j > k\}$. It follows jointly from our choice of i and the exclusion of 0 from E that e_{i-1} is itself excluded from E. Direct

evaluation of the incentive compatibility constraints for S then implies

$$\langle w, P_i \rangle > \frac{e_i - e_{i-1}}{\mu_i - \mu_{i-1}}(\mu_i - \underline{y}) = t_i.$$

In turn $\Pi(W) \leq \pi_i - \langle w, P_i \rangle < \pi_i - \langle w_i, P_i \rangle = \Pi_i = \Pi(W_i)$, where the second equality follows from Theorem 1. The case in which E contains e_j with $j < k$ is symmetric. □

Proof of Theorem 3. Per Lemma 6, S is plausible. Let $\Pi_i = \max(\Pi_1, ..., \Pi_m)$. First we have $\max_W \Pi(W|S) \leq \Pi_i = \Pi(W_i) \leq \Pi(W_i|S)$, where the first inequality follows from Lemma 7, the equality follows from Theorem 1, and the second inequality is vacuous. Accordingly, both inequalities hold with equality and $\max_W \Pi(W|S) = \Pi(W_i|S)$. Second we have $\min_A \Pi(W_i|A) = \Pi_i \geq \Pi(W_i|S)$, where the equality follows again from Theorem 1 and the inequality again from Lemma 7. The inequality again holds with equality and $\min_A \Pi(W_i|A) = \Pi(W_i|S)$. Thus, (W_i, S) is a saddle point. Uniqueness follows from Lemma 8. □

Proof of Proposition 2. Let index i satisfy $\Pi_i = \max(\Pi_1, ..., \Pi_m)$. For any random menu $\nu = (N; (\mathcal{W}_1, c_1), ..., (\mathcal{W}_N, c_N); (\nu_1, ..., \nu_N))$ we have

$$\min_{A \in \mathcal{A}} \Pi(\nu|A) \leq \Pi(\nu|S) \leq \max_j \Pi((\mathcal{W}_j(S), c_j(S))|S) \leq \Pi_i,$$

where the first two inequalities are vacuous and the third follows from Theorem 3. □

C Section 4

Lemma 9. *Let Assumptions 1*, 2* hold. Belief S^* is both well defined and plausible.*

Proof. We retrace the argument given in the proof of Lemma 6, with some superficial adjustments to account for the presence of the multiplicative constant κ^* and with some intermediate mechanical steps left to the reader. Proceeding, we claim first that S^* is composed of well defined probability distributions. Direct adaptation of our

earlier argument yields

$$\sum_{i<l}\tau_i^* \le \kappa^* \sum_{i<l}\frac{a_{i+1}^*-a_i^*}{\overline{y}-a_i^*} \le \kappa^*\frac{a_l^*-\underline{y}}{a_n^*-\underline{y}}\frac{a_n^*-a_1^*}{\overline{y}-a_n^*},$$
$$\sum_{i>l}\tau_i^* \le \kappa^* \sum_{i>l}\frac{a_i^*-a_{i-1}^*}{a_i^*-\underline{y}} \le \kappa^*\frac{\overline{y}-a_l^*}{\overline{y}-a_1^*}\frac{a_n^*-a_1^*}{a_1^*-\underline{y}}.$$

As before, define $\Delta^* \equiv a_n^*-a_1^*$, $\varphi^* \equiv (1/2)(\sqrt{4(\kappa^*)^2+1}+2\kappa^*-1)$, $\underline{\varepsilon}^* \equiv (a_1^*-\underline{y})-\varphi^*\Delta^*$, $\overline{\varepsilon}^* \equiv (\overline{y}-a_n^*)-\varphi^*\Delta^*$ and note that (i) $\underline{\varepsilon}^*, \overline{\varepsilon}^* \ge 0$ per Assumption 2*; (ii) $1+2\varphi^* = (\kappa^*)^{-1}\varphi^*(1+\varphi^*)$ per our choice of φ^*; and (iii) $(\kappa^*)^{-1}\varphi^* > 1$ because $\kappa^* > 0$. We have

$$\kappa^*\frac{a_l^*-\underline{y}}{a_n^*-\underline{y}}\frac{a_n^*-a_1^*}{\overline{y}-a_n^*} = \frac{\kappa^*(a_l^*-\underline{y})\Delta^*}{(\Delta^*+\varphi^*\Delta^*+\underline{\varepsilon}^*)(\varphi^*\Delta^*+\overline{\varepsilon}^*)} \le \frac{a_l^*-\underline{y}}{\overline{y}-\underline{y}}, \tag{15}$$
$$\kappa^*\frac{\overline{y}-a_l^*}{\overline{y}-a_1^*}\frac{a_n^*-a_1^*}{a_1^*-\underline{y}} = \frac{\kappa^*(\overline{y}-a_l^*)\Delta^*}{(\varphi^*\Delta^*+\underline{\varepsilon}^*)(\Delta^*+\varphi^*\Delta^*+\overline{\varepsilon}^*)} \le \frac{\overline{y}-a_l^*}{\overline{y}-\underline{y}}. \tag{16}$$

Together, (15)–(16) imply both $\sum_i \tau_i^* < 1$ and $\underline{y} < r^* < \overline{y}$. This is sufficient to ensure that S^* is composed of well defined elements of $\Delta(\mathcal{Y})$, and we are left to verify that S^* has the appropriate mean for each effort. The $i = l$ case is immediate. Otherwise,

$$\langle y, S_i^*\rangle = \sum_{i\le j\le l}\tau_j^*(\langle y, P_j^*\rangle - \overline{y}) + a_l^* = \sum_{i\le j<l}(a_j^*-a_{j+1}^*) + a_l^* = a_i^* \qquad i<l,$$
$$\langle y, S_i^*\rangle = \sum_{l\le j\le i}\tau_j^*(\langle y, P_j^*\rangle - \underline{y}) + a_l^* = \sum_{l<j\le i}(a_j^*-a_{j-1}^*) + a_l^* = a_i^* \qquad i>l.$$

□

Lemma 10. *Let Assumptions 1*, 2* hold. For every $W = (w, c)$ there exists an index i such that $c(S^*) = e_i$ and $\Pi_i^* \ge \Pi(W|S^*)$.*

Proof. Let $W = (w, c)$ be any contract and note that $\langle w, S^*(e)\rangle - e = \sum_i \xi_i^e\langle w, S_i^*\rangle - e < \sum_i \xi_i^e(\langle w, S_i^*\rangle - e_i^*)$ for all efforts e that are not efficient*. Accordingly, $c(S) = e_i^*$ for some index i. If $i = l$ then $\langle w, P_i^*\rangle \ge 0$ follows immediately from limited liability.

Otherwise we have

$$\langle w, P_i^* \rangle \geq \langle w, P_i^* \rangle - w(\overline{y}) \geq \frac{e_i^* - e_{i+1}^*}{a_{i+1}^* - a_i^*}(\overline{y} - p_i^*) = t_i^* \qquad i < l,$$
$$\langle w, P_i^* \rangle \geq \langle w, P_i^* \rangle - w(\underline{y}) \geq \frac{e_i^* - e_{i-1}^*}{a_i^* - a_{i-1}^*}(p_i^* - \underline{y}) = t_i^* \qquad i > l,$$

where the first pair of inequalities follow from limited liability and the second pair from the definition of τ_i^*. In all three cases we have $\Pi_i^* \geq \Pi(W|S^*)$, as claimed. □

Corollary 1. *Let Assumptions 1*, 2* hold and suppose that the agent's beliefs about expected output are known to be a^*. Contract–recommendation pair W_i^* satisfies $\Pi(W_i^*) = \Pi_i^*$.*

Application of Lemma 4 to the specification of the common knowledge of the relationship between effort and expected output environment with $\mu = a^*$ implies that effort e_i^* is uniformly incentive compatible for W_i^*. Corollary 1 then follows by taking expectations.

Lemma 11. *Let Assumptions 1*, 2* hold and suppose that the agent's beliefs about expected output are known to be a^* and the principal's beliefs have full support. If $W = (w, c)$ and w is not a linear contract then there exists an index i such that $\Pi_i^* > \Pi(W)$.*

Proof. The proof follows that of Lemma 8 very closely. Let $W = (w, c)$ be a nonlinear contract. As before, if w has nonzero minimum payment or if $c(A) = 0$ for any plausible belief A, then W is not a solution to the principal's problem. Accordingly, suppose to the contrary that $c(A) \neq 0$ for all plausible A and that w has minimum payment 0. Write $E \equiv \arg\max_{e \in \mathcal{E}} \langle w, S^*(e) \rangle - e$ for the set of payoff-maximizing effort choices for S^* and note that every effort in E is efficient* per Lemma 10. As before, Fact 2 implies $\Pi(W) \leq \min_{e \in E} \pi(e) - \langle w, P(e) \rangle$. Also as before, it must be that $0 \notin E$ and hence that E contains at least one nonzero effort. Suppose $j > l$ for some j with $e_j \in E$ and set $i \equiv \min\{j | e_j \in E, j > l\}$, with the symmetric $j < l$ case left to the reader. By the same argument as in the proof of the earlier Lemma, we have $\Pi(W) \leq \pi_i^* - \langle w, P_i^* \rangle < \Pi_i^* = \Pi(W_i^*)$. □

Proof of Theorem 4. Per Lemma 9, S^* is plausible. Let $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$. We have $\max_W \Pi(W|S^*) \leq \Pi_i^* = \Pi(W_i^*) \leq \Pi(W_i^*|S^*)$, where the first inequality follows from Lemma 10, the equality from Corollary 1, and the second inequality from the definition of $\Pi(\cdot)$. We also have $\min_{A \in \mathcal{A}} \Pi(W_i^*|S^*) = \Pi_i^* \geq \Pi(W_i^*|S^*)$, where the equality follows again from Corollary 1 and the inequality again from Lemma 10. Accordingly, (W_i^*, S^*) is a saddle point. Uniqueness follows from Lemma 11. □

Proof of Theorem 5. Let $\Pi_i^* = \max(\Pi_1^*, ..., \Pi_n^*)$. We claim that $\Pi(W_i^*|A) \geq \Pi_i^*$ for every plausible belief A. Per Theorem 4, this is sufficient to establish the result. Proceeding, we have $l = 1$ per Assumption 3 and hence $w_i^* = \alpha_i(y - \underline{y})$ for some $\alpha_i \geq 0$. In turn, (12) implies for all efforts e, e' and for all $a \in \mathscr{A}$ $\alpha_i(a(\max\{e, e'\}) - a(\min\{e, e'\})) \geq \alpha_i(a^*(\max\{e, e'\}) - a^*(\min\{e, e'\}))$. Consequently,

$$\langle w_i^*, A(e)\rangle - e > \langle w_i^*, A(e_i^*)\rangle - e_i^* \implies e > e_i^*. \tag{17}$$

Per our choice of i, we have $p(e_i^*) - \alpha_i(p(e_i^*) - \underline{y}) \geq p(e_l^*) > \underline{y}$, where the first inequality follows from $\Pi_i^* \geq \Pi_l^*$ and the second from the interiority assumption maintained throughout the paper. It follows that $\alpha_i < 1$. In turn, Assumption 3 implies

$$e > e_i^* \implies (1 - \alpha_i)p(e) > (1 - \alpha_i)p(e_i^*). \tag{18}$$

The claimed results follow jointly from Theorem 4 and (17)–(18). □

D Counterexample

Example 4. The principal's goal is to maximize guaranteed expected net output $y - w(y)$. Output lies in the interval $[0, 320]$ and the agent chooses between low effort at cost 0, medium effort at cost 5, high effort at cost 35, and maximum effort at cost 50. The principal believes that output is 80 with probability 1 under low effort, 160 with probability 1 under medium effort, 240 with probability 1 under high effort, and 256 with probability 1 under maximum effort. The set of plausible agent beliefs is the set of beliefs consistent with common knowledge of the relationship between effort and expected output.

Suppose the principal in Example 4 pays the agent with the linear bonus contract $w(y) \equiv \max((7/8)(y-160), 0)$ and recommends actions that maximize her own payoff. Per Lemma 1, the range of plausible agent payoffs net of effort costs is $[0, 35]$ for low effort; $[0, 70] - 5 = [-5, 65]$ for medium effort; $[70, 105] - 35 = [35, 70]$ for high effort; and $[84, 112] - 50 = [34, 62]$ for maximum effort. Accordingly, $W \equiv (w, c)$ assigns one of the three nonzero efforts to every plausible belief.

With regards to Sion's sufficient conditions[14] for the existence of saddle point, there exists a plausible belief A with net expected payoffs $(0, 60, 40, 55)$, a plausible belief B with net expected payoffs $(0, 40, 60, 55)$, and a plausible belief $C \equiv (1-\lambda) \cdot A + \lambda \cdot B$ with net expected payoffs $(0, 60(1-\lambda) + 40\lambda, 40(1-\lambda) + 60\lambda, 55)$. The principal's payoff satisfies

$$\Pi(W|A) = 160; \qquad \Pi(W|B) = 170; \qquad \Pi(W|C) = \begin{cases} 160 & 0 \leq \lambda < 1/4, \\ 172 & \lambda \in [1/4, 3/4], \\ 170 & 3/4 < \lambda \leq 1 \end{cases}$$

and thus the map $A \mapsto \Pi(W|A)$ is neither lower semicontinuous nor quasiconvex. Accordingly, Sion's minimax theorem can not be used to establish the existence of a saddle point in general specifications of our problem.

[14] For the purposes of assessing the continuity and convexity criteria in Sion's minimax theorem, we give the set of beliefs $\mathcal{A}$ the product topology and the linear structure $((1-\lambda)A + \lambda B)(e) \equiv (1-\lambda)A(e) + \lambda B(e)$.