Optimal Project Design

Daniel Garrett, George Georgiadis, Alex Smolin, and Bálazs Szentes

July 20, 2020

Abstract

This paper considers a moral hazard model with (i) a risk-neutral agent and (ii) agent limited liability. Prior to interacting with the principal, the agent designs the production technology, which is a specification of the agent's cost of generating each output distribution with support in $[0, 1]$. After observing the production technology, the principal offers a payment scheme and then the agent chooses a distribution over outputs. First, we show that there is an optimal design involving only binary distributions on $\{0, 1\}$; that is, the cost of any other distribution is prohibitively high. Then, we characterize the equilibrium technology defined on the binary distributions and show that the equilibrium payoff of both the principal and the agent is $1/e$. A notable feature of the equilibrium is that the principal is indifferent between offering the equilibrium bonus rewarding output one and anything less than that. Finally, the analysis of the model is shown to generalize to the case where the agent is risk averse.

1 Introduction

A central result in contract theory is that agency rents are a key source of economic welfare. When analyzing environments with asymmetric information, most microeconomic models take the determinants of these agency frictions as given. In hidden-information models, for example, the distribution of types, which determines information rents, is typically treated as exogenous. Similarly, in principal-agent problems with hidden actions, the production technology available to the agent, which governs the principal's cost of implementing various actions, is usually part of the model description. However, if an agent's payoff depends on agency frictions, then he is likely to pursue generating these frictions in a way that enhances his payoff. The goal of this paper is to reconsider the standard limited-liability moral hazard problem and understand how an agent might maximize the rents he obtains due to limited liability.

For a potential application where such a technology design problem may arise, consider the problem of an entrepreneur who must seek venture capital backing in order to start a business. Before contracting, the entrepreneur must develop a business plan which specifies the design of products to offer, the markets the business will operate in, the modes of production, etc. It is conceivable that the entrepreneur has at least some flexibility when developing such a plan. If the venture capitalist has strong bargaining power, the entrepreneur benefits from putting forward a business plan that exacerbates the moral hazard problem and generates large agency rents. Even if there were more profitable alternatives, they may not be considered in the contractual negotiations if the venture capitalist is unaware of them.

In the baseline setup of this paper, we consider a risk-neutral agent who can choose a production technology (or "project") before interacting with a principal.[1] A production technology specifies the agent's cost of each output distribution with support in $[0, 1]$. That is, the only restriction on the available projects is that output is uniformly bounded. Such a bound may represent a physical constraint and is normalized to one. After observing the agent's project, the principal offers a wage contract, which is a mapping from output realizations to monetary compensation. We assume that the agent has limited liability and hence the payment must be non-negative. Finally, the agent chooses an output distribution at a cost determined by his first-stage choice.

Our first main result is that the optimal project involves only binary distributions on $\{0, 1\}$.[2] In other words, the cost of all other distributions can be assumed so high that the principal never wants to implement them, and the agent would never choose them irrespective of the payment scheme. This means that the equilibrium project can be thought of as a task which yields a positive payoff only if completed. The production technology specifies the cost of each probability of completion. The principal's wage contract can be viewed as a bonus paid for project completion, with payments set to zero otherwise.

Let us explain the optimality of binary projects. Just like in standard moral hazard problems, the output plays a dual role in our model. On the one hand, it is the principal's revenue, and on the other hand, it is an informative signal about the agent's action, which is used by the principal to incentivize the agent. By the Informativeness Principle, if this signal is made less informative, incentivizing the agent becomes more expensive. The key observation is that each binary distribution with support $\{0, 1\}$ can be viewed as a garbling of a distribution with the same mean. Consider now a transformation of each project so that, if the agent incurs a cost of a distribution, output is distributed according to the binary distribution with the same mean. This means that the agent's cost of inducing a given level

[1] We then extend our analysis to environments where the agent is risk averse.

[2] While there is some multiplicity of optimal projects, we show below that all optimal projects share the same essential attributes, so we write informally of "the" optimal project.

of expected output remains the same in the transformed project but the principal's cost of implementing it goes up. In this sense, such a transformation exacerbates the moral hazard problem. We show how this observation can be used to replace any project with a binary project for which the agent's payoff is at least as high.

Our second main result is a full characterization of the optimal binary project. In this project, the cost of completing the task with probability $1/e$ is zero. That is, even if the agent incurs no cost, project completion can be achieved with probability $1/e$. In equilibrium, the principal offers a bonus which induces the agent to complete the project with probability one. Furthermore, the principal is indifferent between offering this bonus and anything less than that. This indifference condition pins down the cost to the agent of any probability of success between $1/e$ and one. Since the marginal cost of the success probability is less than one and the maximal output is produced surely, the equilibrium is ex-post efficient. That is, given the equilibrium production technology, the allocation is efficient. It turns out that the optimal project yields an equal split of surplus: both the principal and the agent earn payoff $1/e$.

The first-best social surplus in our model is one since projects that can generate output one at no cost are feasible. Of course, the agent does not choose such a project because then the principal could achieve the maximal output without making any payment. To earn rent, the agent designs the technology so that generating high expected output is artificially costly. In fact, since the equilibrium output is one with probability one, the only source of distortion induced by optimal design relates to this cost. This might be considered a form of "cost padding", different from others identified in the literature.[3]

As mentioned above, an identifying feature of the optimal project is that the principal is indifferent between implementing a large range of completion probabilities. Let us explain the economic reasoning behind this feature. The cost of the equilibrium completion probability is determined by the marginal cost of smaller probabilities. For any bonus offered by the principal, the agent's optimal completion probability is the one that sets the marginal cost equal to the bonus. This means that lowering the marginal cost of success probabilities makes it more attractive to the principal to implement them. However, if the principal strictly prefers to implement the equilibrium probability of completion to implementing some smaller probabilities, then the marginal cost at these smaller probabilities can be lowered without affecting the principal's equilibrium choice. Such a modification of the project decreases the agent's total cost of the equilibrium completion probability and thus increases his overall payoff (while the principal is still willing to offer the same bonus that

[3]For example, Averch and Johnson (1962) observe that a regulated firm has incentives to inflate capital costs.

implements the equilibrium completion probability). The agent can improve any project in this way unless the principal is indifferent between implementing any completion probability which has a positive marginal cost.

We demonstrate that our main results remain valid even if the agent is risk averse. In particular, the search for an optimal project can be still restricted to the set of binary projects. Moreover, the optimal binary project is still ex-post efficient, that is, the principal implements completion probability one. Finally, the optimal binary project is still characterized by the requirement that the principal must be indifferent between offering the equilibrium bonus and anything less than that. Of course, the equilibrium payoffs of the principal and the agent are no longer $1/e$ and they depend on the agent's concave utility function.

Finally, we apply our analysis to a problem which may be of independent interest: the characterization of those payoff combinations which can arise in principal-agent models with limited liability for some exogenously given production technology.[4] The equilibrium payoff profile in our model corresponds to the point in this set where the agent's payoff is maximized. Moreover, an intermediate step in our proof is to identify the largest payoff the agent can achieve as a function of a given profit of the principal. The domain of this function is the interval $[0, 1]$ because the principal can always guarantee a nonnegative profit by offering zero wage and she cannot get more than the maximal output. This function is shown to be strictly concave and zero at the boundaries of its domain. We will argue that a payoff profile can be generated by some production technology if and only if it lies weakly below this curve.

Literature.— The limited liability model of moral hazard for a risk-neutral agent is a staple of introductory courses on contract theory, where a restriction to binary output (which emerges endogenously in our setting) is often made for tractability. A classic reference for limited liability moral hazard is Innes (1990), who demonstrates the optimality of simple debt contracts in a model with a continuum of outputs, and documents the distortion in effort implied by limited liability and moral hazard. More recent treatment of moral hazard with limited liability includes Poblete and Spulber (2012) for a model with a continuum of outputs and Ollier and Thomas (2013) for a model with binary output, but complicated by the presence of adverse selection. While these models feature a risk-neutral agent with limited liability, the alternative friction commonly explored is risk aversion. Seminal work for the moral hazard model with a risk-averse agent includes Mirrlees (1976), Holmström (1979), Grossman and Hart (1983) and Rogerson (1985); see Bolton and Dewatripont (2005) and Holmström (2017) for comprehensive treatments. The focus of the above papers is on

[4]This is similar in spirit to Bergemann et al. (2015), who characterize the set of consumer and seller payoffs in a model of third-degree price discrimination. See also Garrett (2020) for a related exercise in a setting with moral hazard and adverse selection.

optimal contract design, taking the agent's available technology as given. Our paper departs from this approach by viewing the production technology as designed by the agent. We are unaware of such a design problem being posed elsewhere.

There is work that seeks to understand how the primitive contractual environment affects payoffs in moral hazard problems. An example of this is the Informativeness Principle introduced by Holmström (1979) and refined for instance by Chaigneau et al. (2019). These papers clarify how additional information about the agent's action can reduce agency costs for the principal.

In a related paper, Condorelli and Szentes (2020) study the problem of optimally generating information rents in the context of a bilateral trade model. Before interacting with the seller, the buyer can choose the distribution of her valuation for the seller's good. This choice is observed by the seller before she makes a take-it-or-leave-it offer. It turns out that the equilibrium distribution generates a unit-elastic demand, that is, it makes the seller indifferent between setting any price on its support. This is reminiscent of our optimal binary project which makes the seller indifferent across a range of bonuses.[5] This similarity might explain that, when the buyer is restricted to choose distributions with support in the interval $[0, 1]$, the equilibrium payoffs of the buyer and the seller are also $1/e$.[6]

Finally, our paper is related to a growing literature that studies sequential mechanism design problems. In particular, Bhaskar, McClellan, and Sadler (2019) studies insurance market regulation. In their model, the regulator first restricts the set of contracts that the firm is permitted to offer, and the firm then offers a menu of contracts to the agent from the permitted set. While this paper focuses on an adverse selection problem, the common idea that is exploited is that the first-stage designer exploits the deviation incentives in the third stage to discipline the principal in the second stage.

2 Model

We consider a game between a principal (she) and an agent (he), which proceeds as follows. In the first stage, the agent chooses a cost function $c : \mathcal{F} \to \mathbb{R}_+ \cup \{+\infty\}$, where $\mathcal{F}$ denotes the set of CDFs with support on $[0, 1]$. We refer to such a function c as a project. Then, after observing c, the principal offers a payment scheme $w : [0, 1] \to \mathbb{R}_+$, which is restricted to be

[5] Such an indifference argument has appeared in the contexts of incentivizing monitoring (Ortner and Chassang (2018)), optimal testing (Perez-Richet and Skreta (2018)) and monopoly pricing in the presence of ambiguity (Bergemann and Schlag (2011)).

[6] Also related, Roesler and Szentes (2017) consider a setting where signals inform an otherwise uninformed buyer of his value, and asks which signal structure yields the highest information rent.

Borel-measurable.[7] Finally, after observing the offered payment scheme, the agent chooses a distribution $F \in \mathcal{F}$, and the output is realized according to F. If the realized output is x then the agent's and principal's payoffs are $w(x) - c(F)$ and $x - w(x)$, respectively. Both parties are expected payoff maximizers.

Notation.— For each $F \in \mathcal{F}$, let μ_F denote the expected value of F, that is, $\mu_F = \int_0^1 x dF(x)$. The set of projects and the set of Borel-measurable payment schemes are denoted by $\mathcal{C}$ and $\mathcal{W}$, respectively. We refer to a triple $(c, w, F) \in \mathcal{C} \times \mathcal{W} \times \mathcal{F}$ as an outcome. Let U and Π denote the expected payoffs of the agent and the principal defined on the outcomes; that is,

$$U(c, w, F) = \int_0^1 w(x) dF(x) - c(F), \text{ and } \Pi(w, F) = \int_0^1 [x - w(x)] dF(x).$$

Optimal Projects.— In what follows, we take the usual contract-theory approach and wish to define the agent's optimal project as a solution to a maximization problem subject to incentive compatibility. That is, the payment schedule $w \in \mathcal{W}$ that the principal offers should be a best response given the project $c \in \mathcal{C}$, while the distribution $F \in \mathcal{F}$ that the agent chooses should be a best response given the project c and the payment schedule w. We call such a pair (w, F) an *equilibrium* in project c. A formal statement of the incentive compatibility constraints for a best response (for the principal and agent respectively) is given below.

Given a notion of equilibrium, we can then define a project c to be *optimal* if there is an equilibrium (w, F) in c such that the agent's payoff in outcome (c, w, F) is larger than in any outcome (c', w', F') such that (w', F') is an equilibrium in c'. Thus, we assess the optimality of a project c that induces multiple equilibria by considering those which give the highest payoff to the agent. This is in line with the approach prevalent in mechanism design, where the designer is permitted to pick the most favorable equilibrium.

Inventive Compatibility.— We say that choosing F is incentive compatible for the agent in the subgame (c, w) if

$$U(c, w, F) \geq U(c, w, F') \text{ for all } F' \in \mathcal{F}. \tag{1}$$

To describe the principal's incentive constraint is harder because the agent may not have a best response in a subgame generated by a pair (c, w). In turn, this can make it difficult to assess the profitability of certain deviations. To circumvent this problem, we define the *value*

[7] Non-negativity of payments encodes the limited-liability constraint.

of the agent, $u(c,w)$, in each subgame (c,w), by

$$u(c,w) \equiv \sup_{F \in \mathcal{F}} U(c,w,F).$$

We aim to define the *value of the principal* in a subgame (c,w) by reference to sequences of distributions along which the agent's payoff converges to his value. In general, there may be many such sequences, potentially generating different limit payoffs to the principal. Let $\mathbf{F}^{c,w}$ denote the set of sequences of distributions (F_n) along which the agent's payoff converges to $u(c,w)$. Formally, $(F_n) \in \mathbf{F}^{c,w}$ if and only if $\lim_{n\to\infty} U(c,w,F_n) = u(c,w)$. Then, the principal's value in subgame (c,w) is given as

$$\pi(c,w) \equiv \sup \left\{ \limsup_{n\to\infty} \Pi(w,F_n) \quad : \quad (F_n) \in \mathbf{F}^{c,w} \right\}.$$

Evaluating the principal's value by reference to the supremum is again in the spirit of the approach prevalent in mechanism design, where the principal is permitted to pick the most favorable best response of the agent. The principal's incentive compatibility constraint guaranteeing that she offers payment schedule w in project c can be stated as follows: for all $w' \in \mathcal{W}$,

$$\pi(c,w) \geq \pi(c,w').$$

That is, the principal cannot gain by deviating to a payment schedule w', whether or not the agent has a best response to w'.

Binary Projects and Linear Contracts.— As mentioned in the Introduction, binary projects play an important role in our analysis. Next, we formally define these projects. We call a distribution in $\mathcal{F}$ *binary* if its support is in $\{0,1\}$. For each $\mu \in [0,1]$, let B_μ denote the binary CDF which specifies an atom of size μ at one; that is, $B_\mu(x) = (1-\mu) + \mu \mathbb{I}_{\{x=1\}}$. Note that the mean of B_μ is also μ. Let $\mathcal{B}$ denote the set of binary distributions; that is, $\mathcal{B} = \{B_\mu : \mu \in [0,1]\}$. We call a project c *binary* if $c(F) = +\infty$ whenever $F \notin \mathcal{B}$.

In each binary project, the principal always finds it optimal to offer a compensation scheme which pays zero if the output is zero. So, the optimal payment scheme can be summarized by a single bonus, b, which is paid to the agent if the output is one. If the project is binary, the wage at output $x \notin \{0,1\}$ is irrelevant, so such a wage contract can be assumed to be *linear*, denoted by w_b, so that $w_b(x) = bx$ for all x.

If the principal offers a linear contract, w_b, the output distribution affects the payoffs of the agent and the principal only through its mean. That is, whether or not the project is binary,

$$U(c,w_b,F) = \mu_F b - c(F), \text{ and } \Pi(w,F) = \mu_F(1-b). \tag{2}$$

This means that, if the principal offers a linear contract to which the agent has a best response, this best response must involve a distribution which is the least costly among those with the same mean.

3 Main Results

This section is devoted to our two main results. In the next section, we show that it suffices to restrict attention to binary projects and, in Section 3.2, we fully characterize the optimal binary project.

3.1 Binary Projects

In this section, we fix a project c^* and an equilibrium (w^*, F^*) in c^*. Our aim is to construct a binary project $\widetilde{c}$ and an equilibrium $(\widetilde{w}, \widetilde{F})$ in $\widetilde{c}$ so that the outcome $(\widetilde{c}, \widetilde{w}, \widetilde{F})$ Pareto dominates the outcome (c^*, w^*, F^*). Since c^* can be an optimal project, this result implies that there exists an optimal project in the class of binary projects.

As explained in the Introduction, the key observation for this result is that an output realization not only determines the principal's payoff, but also serves as an informative signal about the agent's action. If this signal is made less informative in the sense of Blackwell, incentivizing the agent becomes harder for the principal. To see how an output distribution can be made less informative, consider the following garbling: instead of observing output x, the principal observes output one with probability x and output zero otherwise. That is, the garbling of each $F \in \mathcal{F}$ is B_{μ_F}, so the expected output is unaffected. In fact, B_{μ_F} is the least informative garbling of F, as the same transformation can be applied to any other garbling of F which would again result in B_{μ_F}. So, if the principal could contract only on the realization of B_{μ_F} but not on that of F, her wage cost of implementing F would increase. We next explain how this observation can be used to transform the project c^* to a binary one which is more beneficial for the agent.

The idea behind the construction of the binary project, $\widetilde{c}$, is as follows. We first define the agent's cost of a binary distribution to be the cost of the cheapest distribution in project c^* with the same mean. In this binary project, the principal's wage cost of attaining any level of expected output is higher than in project c^*. In fact, the wage cost of generating μ_{F^*} may be so high that the principal prefers to implement a distribution with a lower mean, thus saving on payments to the agent. In this case, the payoffs of both parties can be lower. Therefore, we further modify the binary project by reducing the agent's cost of $B_{\mu_{F^*}}$ so that the principal can implement it at exactly the same wage cost as that of F^* in c^*.

Before stating the main result of this section, let us introduce an additional piece of notation. Note that the expected payment in outcome (c^*, w^*, F^*) is $\mathbb{E}_{F^*}[w^*]$. If $\mu_{F^*} > 0$ (as must be the case if the outcome is optimal for the agent), we can define b^* to equal $\mathbb{E}_{F^*}[w^*]/\mu_{F^*}$. We can then observe that

$$\mathbb{E}_{F^*}[w^*] = \mu_{F^*} b^* = \mathbb{E}_{F^*}[w_{b^*}] = \mathbb{E}_{B_{\mu_{F^*}}}[w_{b^*}]. \tag{3}$$

That is, the expected payment induced by the pair (w^*, F^*) is the same as that induced by $(w_{b^*}, B_{\mu_{F^*}})$.[8]

Proposition 1. Suppose that (w^*, F^*) is an equilibrium in project c^* with $\mu_{F^*} > 0$. Then there exists a binary project, $\widetilde{c}$, such that

(i) $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in $\widetilde{c}$,

(ii) $U(c^*, w^*, F^*) \leq U(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$, and

(iii) $\Pi(w^*, F^*) = \Pi(w_{b^*}, B_{\mu_{F^*}})$.

Let us describe the binary project $\widetilde{c}$ and the main arguments in the proof of the proposition. It turns out that, in this binary project, the agent's rent can be ensured by making it hard for the principal to dissuade the agent from deviating downwards (i.e., to distributions with lower means). Upwards deviations need not play a role, so we specify the agent's cost of each B_μ with $\mu > \mu_{F^*}$ to be infinity throughout the construction. We now explain the two steps of constructing $\widetilde{c}$ from c^* in more detail. In the first step, we take the agent's cost of $B_{\mu_{F^*}}$ to also be infinite. For each $\mu < \mu_{F^*}$, we specify the cost of B_μ to be the cost of the cheapest distribution in project c^* with expectation μ.[9] We then prove that in order to achieve any expected output, the principal must make a higher expected payment in this binary project than in c^*. In the second step, we redefine the agent's cost of $B_{\mu_{F^*}}$ so that the principal's wage cost of implementing $B_{\mu_{F^*}}$ is exactly $\mathbb{E}_{F^*}[w^*] = \mathbb{E}_{B_{\mu_{F^*}}}[w_{b^*}]$, thus obtaining the project $\widetilde{c}$. We show that the agent's cost of $B_{\mu_{F^*}}$ in $\widetilde{c}$ is less than $c^*(F^*)$. This means that, by Equation (3), Parts (ii) and (iii) of the proposition are satisfied.

We now explain how to obtain Part (i). Note that, by our choice of the agent's cost of the distribution $B_{\mu_{F^*}}$, the agent best responds to w_{b^*} by choosing $B_{\mu_{F^*}}$. Therefore, to prove that $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in project $\widetilde{c}$, we need to demonstrate only that offering w_{b^*} is incentive compatible for the principal. By construction, if the principal wants to implement $B_{\mu_{F^*}}$, she offers payment schedule w_{b^*}. She therefore receives a payoff of $\Pi(w^*, F^*)$. As explained, attaining any other expected output μ ($\mu \neq \mu_{F^*}$) is more expensive for the principal in $\widetilde{c}$ than in c^*. Therefore, since the principal found it optimal to implement

[8] Recall that $w_{b^*}(x) = b^* x$ for all x.

[9] We take the infimum in case no cheapest distribution exists.

F^* in c^*, she optimally chooses to implement $B_{\mu_{F^*}}$ in $\widetilde{c}$ by offering w_{b^*}; that is, w_{b^*} is incentive compatible in $\widetilde{c}$.

Towards the first step described above, let us define the binary project, $\widehat{c}$, as follows:

$$\widehat{c}(B_\mu) = \begin{cases} \inf\{c^*(F) : \mu_F = \mu\} & \text{if } \mu < \mu_{F^*}, \\ \infty & \text{otherwise.} \end{cases}$$

We next formalize the aforementioned implication of the Informativeness Principle; in particular, we demonstrate that the principal is worse off in $\widehat{c}$ than in c^*. In fact, we show that, from the principal's point of view, the transformed project $\widehat{c}$ is worse than being restricted to linear contracts in c^* in the sense that each contract w_b generates weakly more profit to the principal in project c^* than in $\widehat{c}$.

Lemma 1. *For all $b \in [0,1]$, $\pi(\widehat{c}, w_b) \leq \pi(c^*, w_b)$.*

Proof. See the Appendix.

Let us illustrate the argument behind the proof of this lemma for the case where the principal's value in subgame (c^*, w_b), $\pi(c^*, w_b)$, is generated by a best response of the agent. Since the agent's expected payment generated by the linear contract w_b depends only on the expected output, he chooses a distribution only if it is the cheapest among those with the same mean. Therefore, the agent's value in subgame (c^*, w_b) is

$$\sup_{\mu \in [0,1]} \{\mu b - \inf\{c^*(F) : F \in \mathcal{F},\, \mu_F = \mu\}\}. \tag{4}$$

This is the same problem as the one which determines the agent's value in the subgame $(\widehat{c}, w_b)$, except that in the latter, the domain is restricted to be $[0, \mu_{F^*})$. Suppose now that F is incentive compatible in (c^*, w_b), and that μ_F attains the supremum in Problem (4). If $\mu_F < \mu_{F^*}$ then μ_F also solves the agent's problem with the restricted domain, implying that B_{μ_F} is incentive compatible in $(\widehat{c}, w_b)$. In this case, $\pi(\widehat{c}, w_b) = \mu_F(1-b) = \pi(c^*, w_b)$. If $\mu_F \geq \mu_{F^*}$, then the principal's value is at least $\mu_{F^*}(1-b)$ in the subgame (c^*, w_b), so $\pi(c^*, w_b) \geq \mu_{F^*}(1-b) \geq \pi(\widehat{c}, w_b)$, where the second inequality holds because, in project $\widehat{c}$, the agent never chooses a distribution which has mean larger than μ_{F^*}.

We are now ready to define project $\widetilde{c}$. Our aim is to modify $\widehat{c}$ at $B_{\mu_{F^*}}$ so that $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in project $\widetilde{c}$. On the one hand, this requires the cost of $B_{\mu_{F^*}}$ to be sufficiently small to guarantee that $B_{\mu_{F^*}}$ is a best response to w_{b^*}. On the other hand, this cost cannot be too small, for otherwise $B_{\mu_{F^*}}$ could be implemented with a bonus smaller that b^*. Therefore, we specify the cost of $B_{\mu_{F^*}}$ to be the largest cost at which the agent still best responds to

w_{b^*} by choosing $B_{\mu_{F^*}}$. This cost, denoted by $\overline{c}$, satisfies $\mu_{F^*}b^* - \overline{c} = \sup\{\mu b^* - \widehat{c}(B_\mu)\}$. The binary project $\widetilde{c}$ is defined as follows:

$$\widetilde{c}(F) = \begin{cases} \overline{c} & \text{if } F = B_{\mu_{F^*}}, \\ \widehat{c}(F) & \text{if } F \neq B_{\mu_{F^*}}. \end{cases}$$

Next, we demonstrate that the outcome $(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$ Pareto dominates (c^*, w^*, F^*). To this end, we first argue that the cost of $B_{\mu_{F^*}}$ in project $\widetilde{c}$ is weakly smaller than $c^*(F^*)$. Suppose, for a contradiction, that $\overline{c} > c^*(F^*)$. Then,

$$\mu_{F^*}b^* - c^*(F^*) > \mu_{F^*}b^* - \overline{c} = \sup\{\mu b^* - \widehat{c}(B_\mu)\} = \sup\{\mu_F b^* - c^*(F) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\},$$

where the two equalities follow from the definitions of $\overline{c}$ and $\widehat{c}$, respectively. By continuity, this chain implies the existence of $b < b^*$ such that

$$\mu_{F^*}b - c^*(F^*) > \sup\{\mu_F b - c^*(F) : F \in \mathcal{F}, \mu_F < \mu_{F^*}\}.$$

This means that offering the linear contract w_b in project c^* provides the principal with a value at least $\mu_{F^*}(1-b)$, which is strictly more than her equilibrium payoff, $\Pi(w^*, F^*) = \mu_{F^*}(1-b^*)$. This contradicts the incentive compatibility of w^* in c^*.

We are now ready to show that the agent is weakly better off in the outcome $(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$ than in (c^*, w^*, F^*). Indeed,

$$U(c^*, w^*, F^*) = \mu_{F^*}b^* - c^*(F^*) \leq \mu_{F^*}b^* - \widetilde{c}(B_{\mu_{F^*}}) = U(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}}), \tag{5}$$

where the equalities follow from (3) and the inequality follows from $\widetilde{c}(B_{\mu_{F^*}}) = \overline{c} \leq c^*(F^*)$. Also note that (3) implies that the principal's payoffs are the same in these two outcomes,

$$\Pi(w^*, F^*) = \mu_{F^*} - \mathbb{E}_{F^*}[w^*] = \mu_{F^*}(1-b^*) = \Pi(w_{b^*}, B_{\mu_{F^*}}). \tag{6}$$

We defined $\widetilde{c}(B_{\mu_{F^*}})$ so that the payment schedule w_{b^*} implements $B_{\mu_{F^*}}$ in project $\widetilde{c}$. Next, we show that the principal cannot implement $B_{\mu_{F^*}}$ in project $\widetilde{c}$ with any payment schedule w_b such that $b < b^*$. This means that the principal's value from offering w_b in project $\widetilde{c}$ is the same as in project $\widehat{c}$.

Lemma 2. *For all $b \in [0, b^*)$, $\pi(\widetilde{c}, w_b) = \pi(\widehat{c}, w_b)$.*

Proof. See the Appendix.

Let us illustrate the argument of the proof for the case where the agent has a best response

to w_{b^*} in project $\widehat{c}$, say $B_{\mu'}$ (with $\mu' < \mu_{F^*}$); that is, $\mu' b^* - \widehat{c}(B_{\mu'}) = \sup\{\mu b^* - \widehat{c}(B_\mu)\}$. By the definition of $\overline{c}$ and $\widetilde{c}$, this means that $\mu' b^* - \widetilde{c}(B_{\mu'}) = \mu_{F_{\mu^*}} b^* - \widetilde{c}(B_{\mu_{F^*}})$. Since $\mu' < \mu_{F^*}$, this equality implies that for all $b \in [0, b^*)$, $\mu' b - \widetilde{c}(B_{\mu'}) > \mu_{F_{\mu^*}} b - \widetilde{c}(B_{\mu_{F^*}})$, implying that $B_{\mu_{F^*}}$ is not incentive compatible in $(\widetilde{c}, w_b)$. Since the projects $\widetilde{c}$ and $\widehat{c}$ are identical on the rest of their domains, the statement of the lemma follows.

Finally, we are ready to prove Proposition 1.

Proof of Proposition 1. Observe that the outcome $(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$ satisfies Parts (ii) and (iii) of the proposition by Equations (5) and (6), respectively. Therefore, we only need to establish Part (i); that is, we need to argue that $(w_{b^*}, B_{\mu_{F^*}})$ is an equilibrium in project $\widetilde{c}$. By the definition of $\overline{c}$, $B_{\mu_{F^*}}$ is incentive compatible in the subgame $(\widetilde{c}, w_{b^*})$. Next, we prove that w_{b^*} is incentive compatible in $\widetilde{c}$.

If $b > b^*$ then

$$\pi(\widetilde{c}, w_b) \leq \mu_{F^*}(1 - b) < \mu_{F^*}(1 - b^*) = \Pi(w_{b^*}, B_{\mu_{F^*}}),$$

where the first inequality follows from $\widetilde{c}(B_\mu) = \infty$ for all $\mu > \mu_{F^*}$, and the second inequality is implied by $b > b^*$.

If $b < b^*$ then

$$\pi(\widetilde{c}, w_b) \leq \pi(c^*, w_b) \leq \Pi(w^*, F^*) = \Pi(w_{b^*}, B_{\mu_{F^*}})$$

where the first inequality follows from Lemmas 1 and 2, the second one follows from (w^*, F^*) being an equilibrium outcome in project c^*, and the equality is implied by the definition of b^*. **QED**

3.2 Optimal Project

Each binary project, c, can be described by specifying the cost of each probability of success through a function $C : [0, 1] \to \mathbb{R}_+$. In particular, we set $C(\mu) = c(B_\mu)$ for all μ, and we keep in mind that the cost of a non-binary distribution is infinity. Recall that, in binary projects, it is without loss of generality to restrict attention to bonus contracts, w_b, where the agent is paid b if the output is one. Finally, the agent's choice of distribution B_μ can be identified by its mean, μ. In what follows, we describe each binary outcome (c, w_b, B_μ) as (C, b, μ).

We are ready to state the main result of this section.

Proposition 2. There is an optimal binary project, C^*, and an agent-optimal equilibrium in C^*, (b^*, μ^*), such that

(i) $C^{*\prime}(\mu) = 1 - 1/(e\mu)$ if $\mu \geq 1/e$ and zero otherwise,
(ii) $b^* = 1 - 1/e$, and
(iii) $\mu^* = 1$.

Proposition 2 describes an optimal binary project in terms of the marginal costs of completion probabilities. Of course, adding a fix cost has no impact on incentives, so $C^*(0) = 0$. We explain that the functional form of the marginal cost in Part (i) is pinned down by the requirement that the principal must be indifferent between implementing a large range of completion probabilities. We sketch an argument for arriving at Proposition 2 while restricting the agent to choose cost functions C that are non-decreasing, differentiable, and such that equilibrium can be determined using the first-order approach. The formal proof in the Appendix uses an envelope type argument to determine the agent's payoff and does not rely on any such restrictions.

Our first aim is to express the agent's problem of finding an optimal binary project as a maximization problem subject to incentive compatibility constraints. To this end, we first describe the agent's incentive constraint in a subgame (C, b). Note that at this stage the agent's problem is to solve $\max_{\widehat{\mu}\in[0,1]} \{\widehat{\mu}b - C(\widehat{\mu})\}$. Then the requirement on the reward b ensuring the agent chooses a given completion probability μ can be described by the first-order condition

$$b = C'(\mu). \tag{7}$$

We now turn to the incentive constraint of the principal. In project C, the principal's problem is

$$\max_{\mu\in[0,1], b\in\mathbb{R}_+} \mu(1-b)$$

subject to the constraint that μ and b satisfy Equation (7). Plugging the constraint into the maximand, the principal's problem can be expressed solely in terms of the probability of success she wants to implement; that is, $\max_{\mu\in[0,1]} \mu(1 - C'(\mu))$. So, in project C, the principal's choice of μ must satisfy

$$\mu(1 - C'(\mu)) \geq \widetilde{\mu}(1 - C'(\widetilde{\mu})), \tag{8}$$

for all $\widetilde{\mu} \in [0, 1]$.

We are now ready to describe the agent's problem of designing an optimal project. In each project, the principal may be indifferent between implementing various completion probabilities generating different payoffs to the agent. Therefore, we include expected output, μ, in the agent's first stage problem with the interpretation that, after designing a project, the

agent also makes a recommendation to the principal regarding which μ to implement. Then the agent's problem becomes

$$\begin{aligned} &\max_{C,\mu} \mu b - C(\mu) \\ &s.t.\ (7) \text{ and } (8). \end{aligned}$$

Again, plugging the constraint (7) into the maximand, this problem can be rewritten as

$$\begin{aligned} &\max_{C,\mu} \mu C'(\mu) - C(\mu) \\ &s.t.\ (8). \end{aligned} \tag{9}$$

Now let us explain that, if $(\widehat{C}, \widehat{\mu})$ solves this problem, then the constraint (8) binds at each $\widetilde{\mu} < \widehat{\mu}$ if $\widehat{C}'(\widetilde{\mu}) > 0$. The key observation to this is that the agent's cost of $\widehat{\mu}$, $\widehat{C}(\widehat{\mu}) = \int_0^{\widehat{\mu}} \widehat{C}'(\widetilde{\mu})\, d\widetilde{\mu}$,[10] is decreasing in the marginal cost of success at any $\widetilde{\mu} < \widehat{\mu}$, $\widehat{C}'(\widetilde{\mu})$. It is therefore optimal to set these marginal costs in the project C as low as possible. However, they are subject to the lower bound given in the constraint (8). Hence, in the optimum $(\widehat{C}, \widehat{\mu})$, the constraint (8) evaluated at $(C, \mu) = (\widehat{C}, \widehat{\mu})$ binds for all $\widetilde{\mu} < \widehat{\mu}$ as long as $\widehat{C}'(\widetilde{\mu}) > 0$, implying that the principal is indifferent between implementing any expected level of output below $\widehat{\mu}$ which has a strictly positive marginal cost. We can use this observation to express $\widehat{C}$ in terms of the principal's equilibrium payoff, $\widehat{\pi} = \widehat{\mu}(1 - \widehat{C}'(\widehat{\mu}))$. In particular, we have

$$\widehat{C}'(\widetilde{\mu}) = \begin{cases} 1 - \frac{\widehat{\pi}}{\widetilde{\mu}} & \text{if } \widetilde{\mu} \geq \widehat{\pi}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Therefore, given $\widehat{\mu}$ and $\widehat{\pi}$, the principal's binding incentive constraint identifies the optimal project. Consequently, the agent's problem in (9) can be rewritten as

$$\max_{\widehat{\mu}, \widehat{\pi} \in [0,1]} \widehat{\mu}\left(1 - \frac{\widehat{\pi}}{\widehat{\mu}}\right) - \int_{\widehat{\pi}}^{\widehat{\mu}} \left(1 - \frac{\widehat{\pi}}{\widetilde{\mu}}\right) d\widetilde{\mu}. \tag{11}$$

This reformulation of the agent's problem is rigorously justified in the proof of the Appendix, where we consider all binary projects C. Observe that

$$\widehat{\mu}\left(1 - \frac{\widehat{\pi}}{\widehat{\mu}}\right) - \int_{\widehat{\pi}}^{\widehat{\mu}} \left(1 - \frac{\widehat{\pi}}{\widetilde{\mu}}\right) d\widetilde{\mu} = \int_{\widehat{\pi}}^{\widehat{\mu}} \frac{\widehat{\pi}}{\widetilde{\mu}} d\widetilde{\mu} = \widehat{\pi}\left[\log \widehat{\mu} - \log \widehat{\pi}\right], \tag{12}$$

[10] The requirement that $\widehat{C}(0) = 0$ is clearly a necessary condition for an optimal project, since if $\widehat{C}(0) > 0$, we could reduce all costs by this amount, keeping the players' incentives unchanged, but increasing the agent's equilibrium payoff.

which is maximized at $\widehat{\mu} = 1$ and $\widehat{\pi} = 1/e$. This explains how we obtain Parts (ii) and (iii) of the proposition. In particular, the principal's profit in an optimal outcome is $\pi^* = \mu^*(1 - b^*) = 1/e$, and since $\mu^* = 1$, we have $b^* = 1 - 1/e$. Finally, note that evaluating the right-hand side of (10) at $\widehat{\pi} = 1/e$, yields Part (i) of the proposition.

Next, we compute the payoffs of the agent and the principal in the outcome (C^*, b^*, μ^*). As mentioned above, the principal's equilibrium payoff is $\mu^*(1 - b^*) = 1/e$. The agent's payoff is pinned down by evaluating (12) at $(\mu^*, \pi^*) = (1, 1/e)$, also yielding $1/e$. Recall that the optimal project C^* is defined by the principal's binding incentive constraint, (10). This means that the principal can also generate her equilibrium payoff by setting any bonus smaller than b^*. In the Appendix, we show that these conclusions are valid for each optimal binary project. This also implies that, in any optimal project, the agent's optimal payoff is determined by an equilibrium which satisfies Parts (ii) and (iii) of the proposition. Next, we state formally these observations (as derived in the Appendix). We abuse notation and write the players' values, u and π, as functions of a binary outcome.

Remark 1. In any optimal binary project C,
(i) $(b^*, \mu^*) = (1 - e, 1)$ is an agent-optimal equilibrium
(ii) for all $b \in [0, 1 - 1/e]$, $u(C, b) = \int_0^b 1/[e(1 - z)]\, dz$, and
(iii) for all $b \in [0, 1 - 1/e]$, $\pi(C, b) = 1/e$.

Let us now explain that Remark 1 implies that optimal binary projects are "close" to being uniquely determined over completion probabilities in $[1/e, 1]$ in the following sense. If C is also an optimal project, then for each $\widehat{\mu} \in [1/e, 1]$ either $C(\widehat{\mu}) = C^*(\widehat{\mu})$ or there exists a sequence (μ_n) with $\mu_n \to \widehat{\mu}$ such that $\lim C(\mu_n) = C^*(\widehat{\mu})$. To see this, suppose that the principal wants to attain mean $\widehat{\mu} \in [1/e, 1]$ by offering a bonus $\widehat{b}$. By Part (iii) of the remark, this bonus must be the same as the bonus implementing $\widehat{\mu}$ in project C^*. Indeed, since the principal must obtain profit $1/e$, $\widehat{b}$ must be $\widehat{b} = 1 - 1/(\widehat{\mu}e)$ in both C and C^*. By Part (ii), the agent's value in subgame $(C, \widehat{b})$ is the same as in subgame $(C^*, \widehat{b})$, that is, $u(C, \widehat{b}) = u(C^*, \widehat{b})$. If $\widehat{\mu}$ is a best response to the bonus $\widehat{b}$ in C, this implies that $\widehat{b}\widehat{\mu} - C(\widehat{\mu}) = \widehat{b}\widehat{\mu} - C^*(\widehat{\mu})$ and consequently, $C(\widehat{\mu}) = C^*(\widehat{\mu})$. If the agent's value is not generated by a best-response in $(C, \widehat{b})$, there must be a sequence (μ_n) converging to $\widehat{\mu}$ such that $\lim\{\widehat{b}\mu_n - C(\mu_n)\} = \widehat{b}\widehat{\mu} - C^*(\widehat{\mu})$ and hence, $\lim C(\mu_n) = C^*(\widehat{\mu})$. Note that if we require that C is continuous, we have that $C(\mu) = C^*(\mu)$ for all $\mu \in [1/e, 1]$. However, no such restrictions apply to the costs of completing the project with "low probabilities", i.e. for probabilities below $1/e$. Since the agent can induce an expected output of $1/e$ at no cost, the completion probabilities below $1/e$ are, in effect, redundant.

4 Discussion

Uniqueness.— The above discussion describes the multiplicity of optimal binary projects, emphasizing that certain key properties of such projects are uniquely determined. Still, the possibility of optimal but non-binary projects may be a further source of non-uniqueness. Nonetheless, we show that properties analogous to the ones stated in Remark 1 continue to hold, even among non-binary projects.

We first explain that the output is one in any agent-optimal equilibrium of any optimal project. To this end, let c^* be an optimal project and (w^*, F^*) be an agent-optimal equilibrium in c^*. Proposition 1 applied to the optimal outcome (c^*, w^*, F^*) implies that the corresponding binary outcome $(\widetilde{c}, w_{b^*}, B_{\mu_{F^*}})$ is also optimal. Then, Part (i) of Remark 1 implies that we must have $\mu_{F^*} = 1$, and hence, $F^* = B_1$.

Next we argue that output realizations in $(0, 1)$ are redundant in the sense that replacing these realizations by zero has no impact on equilibrium behavior. To this end, we first show that it can be assumed that $w^*(x) = 0$ for all $x < 1$. The intuition is that if the principal wants to implement output one, she should not reward the agent for any other output realization by offering a positive payment. To state it formally, let us define the payments scheme, ρ_b, for each b such that $\rho_b(1) = b$ and $\rho_b(x) = 0$ for $x \neq 1$. Observe that replacing w^* by $\rho_{w^*(1)}$ makes choosing B_1 no less attractive to the agent, so $\left(\rho_{w^*(1)}, B_1\right)$ is also an agent-optimal equilibrium in c^*.

Provided that the compensation scheme is $\rho_{w^*(1)}$, when the agent is contemplating choosing a distribution, all that matters is the probability that the output is one. So, moving all the probability mass from $(0, 1)$ to zero has no impact on the agent's choice of a distribution. Moreover, these new distributions generate smaller expected outputs, hence the principal still prefers to implement B_1. To make these claims precise, let us define a binary project $\widetilde{C}$ such that $\widetilde{C}(\mu) = \inf\{c^*(F) : \Delta(F) = \mu\}$, where $\Delta(F)$ denotes the atom at one specified by F.[11] Given the payment schedule $\rho_{w^*(1)}$, the change in cost function from c^* to $\widetilde{C}$ does not affect the agent's willingness to choose completion probability one. We conclude that the binary project $\widetilde{C}$ is optimal and $(w^*(1), 1)$ is an agent-optimal equilibrium in $\widetilde{C}$.

To give a further sense in which the results in the previous section are robust to the considerations of non-binary projects, we state the following

Remark 2. If c^* is an optimal project and (w^*, F^*) is an agent-optimal equilibrium in c^*, then

(i) $F^* = B_1$,

(ii) $\Pi(w^*, F^*) = U(c^*, w^*, F^*) = 1/e$,

[11] That is, $\Delta(F) = F(1) - \lim_{x \nearrow 1} F(x)$.

(iii) $u\left(c^*, \rho_b\right) = \int_0^b 1/\left[e\left(1-z\right)\right] dz$ for all $b \in [0, 1-1/e]$, and
(iv) $\pi\left(c^*, \rho_b\right) = 1/e$ for all $b \in [0, 1-1/e]$.

We have already established Part (i) and that the binary outcome $(\widetilde{C}, w^*\left(1\right), 1)$ is agent-optimal. So, by Part (iii) of Remark 1, $w^*\left(1\right) = 1-1/e$, and the principal's payoff in project c^* must be $1/e$, establishing Part (ii). By the construction of $\widetilde{C}$, $u(c^*, \rho_b)$ is equal to $u(\widetilde{C}, b)$ for all b. Optimality of $\widetilde{C}$ and Part (ii) of Remark 1 then imply Part (iii) of Remark 2. Part (iii) of Remark 1 implies $\pi(\widetilde{C}, b) = 1/e$ for any $b \in [0, 1-1/e]$ and hence $\pi\left(c^*, \rho_b\right) \geq 1/e$. Part (ii) of the remark then implies $\pi\left(c^*, \rho_b\right) = 1/e$ for all $b \in [0, 1-1/e]$ establishing Part (iv). This remark implies that the results stated in Remark 1 remain valid even when projects are not restricted to be binary. If the principal offers payment schedule ρ_b, $b \in [0, 1-1/e]$, then the players' values are the same in the project c^* and in the binary project $\widetilde{C}$.

Payoff Possibility Set.—We can use our results to characterize the set of possible payoff combinations which can arise in principal-agent models for some production technology.[12] To be more specific, we still consider an environment where the agent is risk neutral and has limited liability. However, the production technology is exogenously given and satisfies the constraint that the largest output cannot exceed one. Moreover, the agent has an outside option we take to be zero. We wish to characterize those payoff profiles which can arise in equilibrium for some production technology. We next show that there exists a production technology where the principal receives $\widehat{\pi}$ and the agent receives $\widehat{u}$ if and only if $\widehat{\pi} \in [0, 1]$ and $\widehat{u} \in [0, -\widehat{\pi}\log\widehat{\pi}]$; see Figure 1 for illustration.

First note that the principal's payoff, $\widehat{\pi}$, must be between zero and one. The reason is that the contract w_0 guarantees at least zero profit. Since output is less than one, limited liability implies that the profit cannot exceed one. Next, we identify the frontier of the payoff possibility set. That is, we compute the largest payoff the agent can get if the principal's payoff is $\widehat{\pi}$. By Proposition 1, we know that this largest payoff is achieved in a binary project. Furthermore, recall that in Section 3.2 we rewrote the problem of designing the optimal binary project as a maximization problem with respect to equilibrium probability of success, $\widehat{\mu}$, and the principal's equilibrium profit, $\widehat{\pi}$; see (11). So the problem of designing the agent-optimal binary technology which generates $\widehat{\pi}$ can be reduced to a similar maximization problem except $\widehat{\pi}$ is treated as a parameter instead of a choice variable. By Equation (12), the agent's maximal payoff is $-\widehat{\pi}\log\widehat{\pi}$.

Since the agent's outside option is zero, his equilibrium payoff cannot be negative. It remains to argue that given the principal's payoff, $\widehat{\pi}$, for each $\widehat{u} \in [0, -\widehat{\pi}\log\widehat{\pi})$, there is a

[12]This is similar in spirit to Bergemann et al. (2015), who characterize the set of consumer and seller payoffs in a model of third-degree price discrimination. See also Garrett (2020) for a related exercise in a setting with moral hazard and adverse selection.

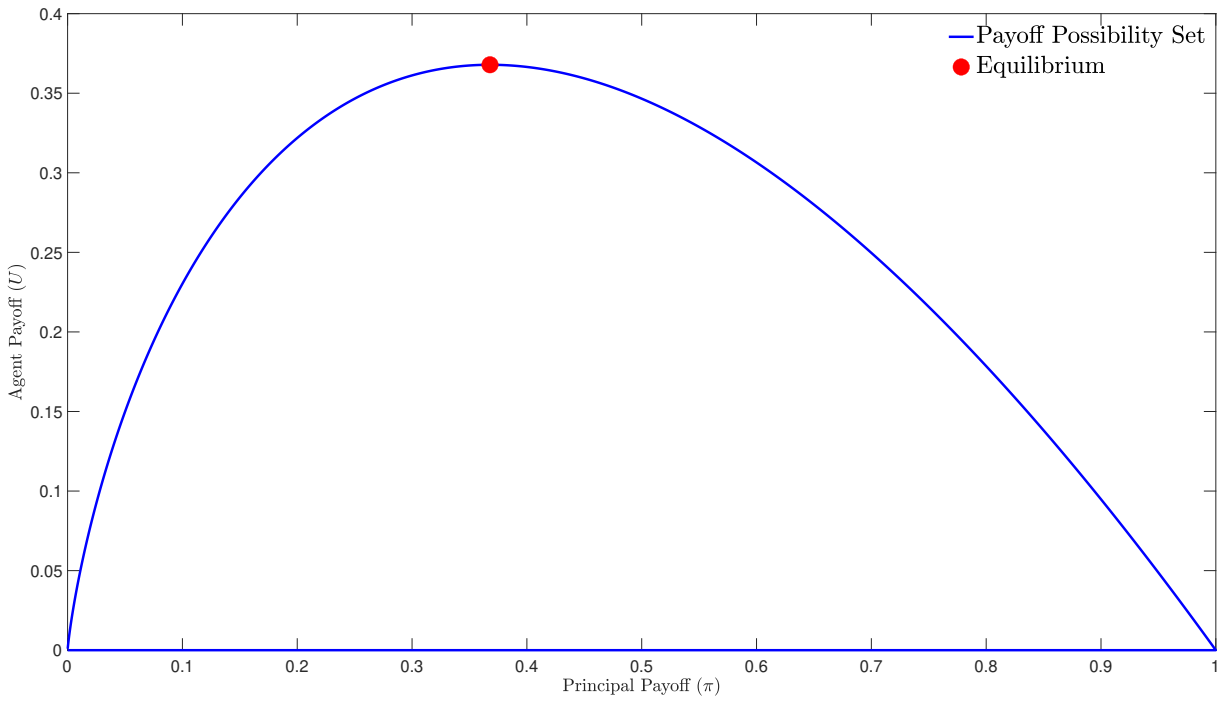

Figure 1: Payoff Possibility Set

production technology so that the equilibrium payoff profile is $(\widehat{\pi}, \widehat{u})$. To do so, consider the production technology generating $(\widehat{\pi}, -\widehat{\pi}\log\widehat{\pi})$ and add a fix cost of $-\widehat{\pi}\log\widehat{\pi} - \widehat{u}$. That is, the agent's cost of each distribution is increased by this quantity. Of course, adding this fixed cost does not change the agent's incentives but lowers his payoff to $\widehat{u}$.

Infinitely Many Distributions.— The optimal project is characterized by a continuum of binding incentive constraints for the principal. In particular, the agent could not achieve the same payoff if the domain of the projects were restricted to be finite. It turns out that, if the agent is restricted to designing projects with finitely many binary distributions, then increasing the number of distributions permitted strictly increases his payoff. We explain this observation by showing how a project with binary domain can be improved by adding an additional distribution to its domain.

Consider first a binary project with only two distributions B_{μ_L} and B_{μ_H}, $\mu_L < \mu_H$. Let's also fix $C(\mu_L)$ and compute the agent-optimal choice of $C(\mu_H)$. Assuming $C(\mu_L) < C(\mu_H)$, it is clear that offering zero bonus implements B_{μ_L}. The incentive constraint for a bonus b_H to induce the agent to take the high action is then

$$\mu_H b_H - C(\mu_H) \geq \mu_L b_H - C(\mu_L),$$

and the principal-optimal bonus ensures this holds as an equality. Also, in projects constrained to have finitely many distributions, one can obtain a similar observation as for the optimal project considered above: the principal must be indifferent between implementing any of the available distributions. In the case of only two binary distributions, this means that

$$\mu_L = \mu_H (1 - b_H),$$

which determines b_H. Note that the left-hand side is the principal's payoff if she offers bonus zero and the agent chooses μ_L. The right-hand side is her payoff is she offers bonus b_H and the agent chooses μ_H. Therefore, the optimal specification of the cost of the high probability is

$$C(\mu_H) = C(\mu_L) + (\mu_H - \mu_L) b_H,$$

with $C(\mu_L)$ exogenously fixed above, and with b_H satisfying the principal's indifference condition.

Now consider how the agent can get the same bonus at a lower cost $C(\mu_H)$ of μ_H if there is a third distribution, B_{μ_M}, such that $\mu_M \in (\mu_L, \mu_H)$. In this case, the bonuses implementing μ_M and μ_H in the optimal project, b_M and b_H, are defined by the principal's indifference conditions

$$\mu_L = \mu_M (1 - b_M) = \mu_H (1 - b_H). \tag{13}$$

Note that the bonus b_H defined by the equality chain is the same as before, which means that b_H is the same as with two distributions.

Now note that incentive compatibility is guaranteed if we can verify downward incentive constraints. Thus, by the same logic as above, we must have

$$C(\mu_M) = C(\mu_L) + (\mu_M - \mu_L) b_M,$$

and

$$\begin{aligned} C(\mu_H) &= C(\mu_M) + (\mu_H - \mu_M) b_H \\ &= C(\mu_L) + (\mu_M - \mu_L) b_M + (\mu_H - \mu_M) b_H, \end{aligned}$$

where b_M and b_H are determined by the indifference conditions in Equation (13). Since $\mu_M < \mu_H$, we have that $b_M < b_H$, which shows that $C(\mu_H)$ decreases as compared to the case with two distributions. Hence, the agent's payoff strictly increases.

Risk Aversion.—Our results can be generalized to the case where the agent is risk averse. To explain it more formally, let us modify our model so that, if the agent receives payment

w and chooses F at cost $c(F)$, then his payoff is $v(w) - c(F)$, where v is an increasing, concave and continuously differentiable function. It can be shown that even in this case, binary projects remain optimal. More precisely, the following version of the statement of Proposition 1 remains valid. For each project c^* and any equilibrium (w^*, F^*) in c^*, there is a binary project $\widetilde{c}$ with an equilibrium $(\widetilde{w}, B_1)$ such that the outcome $(\widetilde{c}, \widetilde{w}, B_1)$ Pareto dominates (c^*, w^*, F^*). The characterization of optimal binary projects follows the same steps as for the risk-neutral case. In this case, the marginal cost of any completion probability $\mu > \pi^*$ is $v(1 - [\pi^*/\mu])$ where π^* is the equilibrium payoff of the principal.

5 Appendix: Omitted proofs

5.1 Proofs of results in Section 3.1

Proof of Lemma 1. Consider a sequence (μ_n), with $\mu_n \in [0, \mu_{F^*})$ for all n, such that

$$u(\widehat{c}, w_b) = \lim_{n\to\infty} U(\widehat{c}, w_b, B_{\mu_n}) \text{ and } \pi(\widehat{c}, w_b) = \lim_{n\to\infty} \Pi(w_b, B_{\mu_n}).$$

For each $k \in \mathbb{N}$, there exists n_k such that, for all $\mu \in [0, 1]$,

$$\mu_{n_k} b - \widehat{c}\left(B_{\mu_{n_k}}\right) + \frac{1}{k} \geq \mu b - \widehat{c}(B_\mu).$$

Equivalently, for all k, and all $\mu \in [0, \mu_{F^*})$,

$$\mu_{n_k} b - \inf\{c^*(F) : \mu_F = \mu_{n_k}\} + \frac{1}{k} \geq \mu b - \inf\{c^*(F) : \mu_F = \mu\}.$$

For each k, we can pick a distribution F_{n_k} with mean μ_{n_k} such that

$$c^*(F_{n_k}) < \inf\{c^*(F) : \mu_F = \mu_{n_k}\} + \frac{1}{k}.$$

Then, for all k and all $F \in \mathcal{F}$ with mean $\mu_F \in [0, \mu_{F^*})$,

$$\mu_{n_k} b - c^*(F_{n_k}) + \frac{2}{k} > \mu_F b - c^*(F). \tag{14}$$

There are then two cases. In the first, the inequality (14) holds for all k and all $F \in \mathcal{F}$ (not only those F with $\mu_F < \mu_{F^*}$). Then

$$u(c^*, w_b) = \lim_{k\to\infty} U(c^*, w_b, F_{n_k})$$

and hence

$$\pi\left(c^{*}, w_{b}\right) \geq \lim_{k \to \infty} \Pi\left(w_{b}, F_{n_{k}}\right) = \lim_{k \to \infty} \Pi\left(w_{b}, B_{\mu_{n_{k}}}\right) = \pi\left(\widehat{c}, w_{b}\right)$$

as desired. In the second, the inequality (14) fails to hold for some k and some $F \in \mathcal{F}$ with $\mu_{F} \geq \mu_{F^{*}}$, which implies

$$u\left(c^{*}, w_{b}\right) = \sup\left\{\mu_{F} b - c^{*}\left(F\right) : F \in \mathcal{F}\right\} > \sup\left\{\mu_{F} b - c^{*}\left(F\right) : F \in \mathcal{F}, \mu_{F} < \mu_{F^{*}}\right\}.$$

This means that there is a sequence of distributions in $\mathcal{F}$ along which the agent's payoff converges to his value $u\left(c^{*}, w_{b}\right)$ and for which every distribution has mean at least $\mu_{F^{*}}$. By the definition of the principal's value, we have

$$\pi\left(c^{*}, w_{b}\right) \geq \mu_{F^{*}}\left(1 - b\right) \geq \pi\left(\widehat{c}, w_{b}\right),$$

where the second inequality follows because any distribution with mean at least $\mu_{F^{*}}$ is assigned an infinite cost in the project $\widehat{c}$.

QED

Proof of Lemma 2. Let us fix $b \in [0, b^{*})$. We first show that w_{b} does not implement $B_{\mu_{F^{*}}}$ in $(\widetilde{c}, w_{b})$. Suppose for a contradiction that $B_{\mu_{F^{*}}}$ satisfies the agent's incentive constraint in $(\widetilde{c}, w_{b})$, that is,

$$\mu_{F^{*}} b - \overline{c} \geq \sup_{\mu < \mu_{F^{*}}} \left\{\mu b - \widetilde{c}\left(B_{\mu}\right)\right\}. \tag{15}$$

Therefore,

$$\overline{c} \leq - \sup_{\mu < \mu_{F^{*}}} \left\{\left(\mu - \mu_{F^{*}}\right) b - \widetilde{c}\left(B_{\mu}\right)\right\} \leq - \sup_{\mu < \mu_{F^{*}}} \left\{\left(\mu - \mu_{F^{*}}\right) b^{*} - \widetilde{c}\left(B_{\mu}\right)\right\} = \overline{c},$$

where the first inequality is just the previous displayed inequality rearranged, the second inequality follows from $b < b^{*}$, and the equality is the definition of $\overline{c}$. Since the farthest left term and the farthest right term are equal in the previous chain, all inequalities must be equalities. Note that the second inequality is an equality only if the supremum in Equation (15) is approached along a sequence of μ's converging to $\mu_{F^{*}}$. Since $\widetilde{c}\left(B_{\mu}\right) = \widehat{c}\left(B_{\mu}\right)$ whenever $\mu \neq \mu_{F^{*}}$, and since $\widehat{c}\left(B_{\mu}\right) = \infty$ for $\mu \geq \mu_{F^{*}}$, it follows that the supremum of $\mu b - \widehat{c}\left(B_{\mu}\right)$ is approached by the same sequence. Hence, $\pi\left(\widehat{c}, w_{b}\right) = \mu_{F^{*}}\left(1 - b\right)$. We can conclude that

$$\pi\left(c^{*}, w_{b}\right) \geq \pi\left(\widehat{c}, w_{b}\right) = \mu_{F^{*}}\left(1 - b\right) > \mu_{F^{*}}\left(1 - b^{*}\right) = \Pi\left(w^{*}, F^{*}\right),$$

where the first inequality follows from Lemma 1, the strict inequality is implied by $b < b^{*}$ and the second equality follows from the definition of b^{*}. This inequality implies that w^{*} is

not incentive compatible in project c^*, a contradiction.

Since w_b does not implement $B_{\mu_{F^*}}$ in $\widetilde{c}$, $U(\widetilde{c}, w_b, B_{\mu_{F^*}}) < u(\widetilde{c}, w_b)$. We next show that $\mathbf{F}^{\widetilde{c},w_b} = \mathbf{F}^{\widehat{c},w_b}$.[13] Note that, for each $(F_n) \in \mathbf{F}^{\widetilde{c},w_b} \cup \mathbf{F}^{\widehat{c},w_b}$, there exists $K \in \mathbb{N}$ such that $F_k \neq B_{\mu_{F^*}}$ if $k > K$. If $(F_n) \in \mathbf{F}^{\widehat{c},w_b}$, it follows from $\widehat{c}(B_{\mu_{F^*}}) = \infty$. If $(F_n) \in \mathbf{F}^{\widetilde{c},w_b}$, it is implied by $U(\widetilde{c}, w_b, B_{\mu_{F^*}}) < u(\widetilde{c}, w_b)$. Since $\widetilde{c}(F) = \widehat{c}(F)$ whenever $F \neq B_{\mu_{F^*}}$, this means that, for each $(F_n) \in \mathbf{F}^{\widetilde{c},w_b} \cup \mathbf{F}^{\widehat{c},w_b}$,

$$\lim_{n\to\infty} U(\widetilde{c}, w_b, F_n) = \lim_{n\to\infty} U(\widehat{c}, w_b, F_n),$$

implying that $\mathbf{F}^{\widetilde{c},w_b} = \mathbf{F}^{\widehat{c},w_b}$. Consequently,

$$\begin{aligned}
\pi(\widetilde{c}, w_b) &\equiv \sup\left\{\limsup_{n\to\infty} \Pi(w_b, F_n) \;:\; (F_n) \in \mathbf{F}^{\widetilde{c},w_b}\right\} \\
&= \sup\left\{\limsup_{n\to\infty} \Pi(w_b, F_n) \;:\; (F_n) \in \mathbf{F}^{\widehat{c},w_b}\right\} = \pi(\widehat{c}, w_b).
\end{aligned}$$

QED

5.2 Proof of Proposition 3.2 and Remark 1

This section proves Proposition 3.2, and in the process Remark 1 (note that Remark 2 is established in the main text). We begin by considering an arbitrary binary project C. Let us drop the dependence on C and write the value for the agent when the bonus is b as $u(b)$. The value is given by

$$u(b) = \sup_{\mu\in[0,1]} \{b\mu - C(\mu)\}.$$

Note that u is non-decreasing. Moreover, as the upper envelope of linear functions, it is convex and hence continuous.

Let $\Gamma(b)$ be the set of values μ such that there is a sequence (μ_n) with $\mu_n \to \mu$ and $b\mu_n - C(\mu_n) \to u(b)$. Take $\bar{\mu}(b) = \max \Gamma(b)$, and note that the maximum is attained. Similarly, let $\underline{\mu}(b)$ be the minimum of $\Gamma(b)$ (also attained). Note that, if the principal offers a bonus $b \in [0,1]$ in project C, then she obtains value $\bar{\mu}(b)(1-b)$.

For any $b \geq 0$, let $u'_+(b)$ be the right derivative of u at b. For any $b > 0$, let $u'_-(b)$ be the left derivative of u at b. We next show a result that is analogous to Theorem 1 of Milgrom and Segal (2002), but adjusted for the possibility that the agent's payoff $u(b)$ is not attained by values $\mu \in \Gamma(b)$.

[13] Recall that $(F_n) \in \mathbf{F}^{c,w}$ if and only if $\lim_{n\to\infty} U(c, w, F_n) = u(c, w)$.

Lemma 3. *For all $b \geq 0$, $u'_+(b) \geq \bar{\mu}(b)$. For all $b > 0$, $u'_-(b) \leq \underline{\mu}(b)$.*

Proof of Lemma 3. Consider a sequence (μ_n) with $\mu_n \to \bar{\mu}(b)$ and $b\mu_n - C(\mu_n) \to u(b)$. Then, for any $b' > b$, we have

$$(b' - b)\,\bar{\mu}(b) = \lim_{n\to\infty} \{b'\mu_n - C(\mu_n)\} - \lim_{n\to\infty} \{b\mu_n - C(\mu_n)\} \leq u(b') - u(b).$$

Dividing by $b' - b$ and taking limits as b' approaches b from above yields $u'_+(b) \geq \bar{\mu}(b)$.

Let $b > 0$ and consider a sequence (μ_n) with $\mu_n \to \underline{\mu}(b)$ and $b\mu_n - C(\mu_n) \to u(b)$. For any $b' < b$, we have

$$(b - b')\,\underline{\mu}(b) = \lim_{n\to\infty} \{b\mu_n - C(\mu_n)\} - \lim_{n\to\infty} \{b'\mu_n - C(\mu_n)\} \geq u(b) - u(b').$$

Dividing by $b - b'$ and taking limits as b' approaches b from below yields $u'_-(b) \leq \underline{\mu}(b)$. **QED**

We can further use the convexity of u to determine its right derivative in terms of the completion probability attainable with a given bonus.

Lemma 4. *For all $b \geq 0$, $u'_+(b) = \bar{\mu}(b)$.*

Proof of Lemma 4. Fix $b \geq 0$ and suppose for a contradiction that $u'_+(b) > \bar{\mu}(b)$. By convexity of u and the previous lemma

$$\bar{\mu}(b) < u'_+(b) \leq u'_-(b') \leq \underline{\mu}(b')$$

for all $b' > b$. For each $n \in \mathbb{N}$, let

$$b_n \in \left(b, b + \frac{1}{n}\right)$$

and let $\mu_n \in \left[\frac{\bar{\mu}(b)+u'_+(b)}{2}, 1\right]$ and such that

$$b_n\mu_n - C(\mu_n) > u(b_n) - \frac{1}{n}$$

(that such a choice is possible follows because $\frac{\bar{\mu}(b)+u'_+(b)}{2} < u'_-(b_n) \leq \underline{\mu}(b_n)$ for all n). Consider a subsequence (b_{n_k}) such that $\mu_{n_k} \to \mu^* \geq \frac{\bar{\mu}(b)+u'_+(b)}{2}$ for some μ^*. We have

$$\lim \{b\mu_{n_k} - C(\mu_{n_k})\} = \lim \{b_{n_k}\mu_{n_k} - C(\mu_{n_k})\} = \lim u(b_{n_k}) = u(b)$$

where the final equality follows by continuity of u. The fact that $\mu^* > \bar{\mu}(b)$ contradicts the definition of $\bar{\mu}(b)$. **QED**

Note now that, because u is convex, it is absolutely continuous and hence differentiable almost everywhere. This means that

$$u(b) = u(0) + \int_0^b \bar{\mu}(s)\, ds.$$

It is immediate from the agent's problem that we must have $u(0) \leq 0$; i.e., the agent cannot obtain a strictly positive payoff if the bonus is set to zero.

Consider now a project C with an equilibrium in which the principal offers bonus $\widehat{b}$ for project completion, the agent chooses completion probability $\widehat{\mu}$, and therefore the principal's payoff is given by $\widehat{\pi} = \widehat{\mu}(1 - \widehat{b})$. Incentive compatibility of the principal offering bonus $\widehat{b}$ requires that, for all b,

$$\begin{aligned} \widehat{\pi} &\geq \bar{\mu}(b)(1-b) \\ &= u'_+(b)(1-b). \end{aligned} \tag{16}$$

Now let us determine the highest agent value, across projects C, that can occur for an equilibrium in which the principal offers bonus $\widehat{b}$ for completion and the agent chooses completion probability $\widehat{\mu}$. Consider then the problem of maximizing the agent's equilibrium payoff

$$u(\widehat{b}) = u(0) + \int_0^{\widehat{b}} u'_+(b)\, db$$

by choice of convex function $u : \mathbb{R}_+ \to \mathbb{R}$ satisfying (i) $u(0) \leq 0$, and (ii) $\widehat{\pi} \geq u'_+(b)(1-b)$ for all b. The first requirement reflects the above observation that the agent cannot obtain a positive payoff if the bonus is zero. The second condition is a re-statement of Condition (16). Any solution to this problem involves $u(0) = 0$ and

$$u'_+(b) = \frac{\widehat{\pi}}{1-b}$$

for all $b \in [0, \widehat{b}]$. In other words, the constraint (ii), or equivalently (16), holds with equality over $b \in [0, \widehat{b}]$ (in which case, the principal must obtain payoff $\widehat{\pi}$ from all such bonuses b). The agent's value function is therefore given by

$$u(b) = \int_0^b \frac{\widehat{\pi}}{1-z} dz \tag{17}$$

on $[0, \widehat{b}]$.

Now, recall that $\widehat{\pi} = \widehat{\mu}(1 - \widehat{b})$, or $\widehat{b} = 1 - \frac{\widehat{\pi}}{\widehat{\mu}}$. The agent's equilibrium payoff can then be

written as

$$\int_0^{1-\frac{\widehat{\pi}}{\widehat{\mu}}} \frac{\widehat{\pi}}{1-z} dz = \left[-\widehat{\pi}\log(1-z)\right]_0^{1-\frac{\widehat{\pi}}{\widehat{\mu}}} = \widehat{\pi}\left(\log(\widehat{\mu}) - \log(\widehat{\pi})\right)$$

which is the expression given in Equation (12) (hence establishing also the one in Equation (11)). As explained in the main text, this payoff is maximized across feasible equilibrium values of $\widehat{\mu}$ and $\widehat{\pi}$ by $\widehat{\mu} = 1$ and $\widehat{\pi} = \frac{1}{e}$. The corresponding equilibrium bonus must be $\widehat{b} = 1 - 1/e$.

Note then that, if the project is C^* as given in the proposition, and the principal offers any $b \in [0, 1-1/e]$, the agent best responds by choosing μ such that

$$b = 1 - \frac{1}{e\mu},$$

i.e. $\mu = \frac{1}{e(1-b)}$. All such bonuses therefore generate profit $1/e$ for the principal. Hence, it is indeed an equilibrium of project C^* for the principal to offer bonus $b^* = 1 - 1/e$, and the agent to choose completion probability equal to $\mu^* = 1$. This completes the proof of the proposition.

Finally, consider Remark 1. For an optimal project, as observed above, the agent-optimal equilibrium delivers the principal profits $\widehat{\pi} = \frac{1}{e}$. Given this, Part (i) of the remark is simply Equation (17). Part (ii) of the remark follows from the above observation that the agent best responds to a bonus $b \in [0, 1-1/e]$ with a binary distribution with mean $\mu = \frac{1}{e(1-b)}$. **QED**

References

Averch, H. and Johnson, L.L., 1962. Behavior of the firm under regulatory constraint. *American Economic Review*, 52(5), pp.1052-1069.

Bergemann, D. and Schlag, K., 2011. Robust monopoly pricing. Journal of Economic Theory, 146(6), pp.2527-2543.

Bergemann, D., Brooks, B. and Morris, S., 2015. The limits of price discrimination. *American Economic Review*, 105(3), pp.921-57.

Bhaskar, B., McClellan, A., and Sadler, E., 2019. Regulation design in insurance markets. Working Paper.

Bolton, P. and Dewatripont, M., 2005. Contract theory. MIT Press.

Chaigneau, P., Edmans, A. and Gottlieb, D., 2019. The informativeness principle without the first-order approach. *Games and Economic Behavior*, 113, pp.743-755.

Condorelli, D. and Szentes, B., Forthcoming. Buyer-optimal demand and monopoly pricing. *Journal of Political Economy*.

Garrett, D., 2020. Payoff Implications of Incentive Contracting. Working Paper.

Grossman, S.J. and Hart, O.D., 1983. An Analysis of the Principal-Agent Problem. *Econometrica*, 51(1), pp.7-46.

Holmström, B., 1979. Moral hazard and observability. *Bell Journal of Economics*, pp.74-91.

Holmström, B., 2017. Pay for performance and beyond. *American Economic Review,* 107(7), pp.1753-77.

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

Milgrom, P. and Segal, I., 2002. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2), pp.583-601.

Mirrlees, J.A., 1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics*, pp. 105-131.

Ollier, S. and Thomas, L., 2013. Ex post participation constraint in a principal–agent model with adverse selection and moral hazard. *Journal of Economic Theory*, 148(6), pp.2383-2403.

Ortner, J. and Chassang, S., 2018. Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy*, 126(5), pp.2108-2133.

Perez-Richet, E. and Skreta, V., 2018. Test Design Under Falsification. Working Paper.

Poblete, J. and Spulber, D., 2012. The form of incentive contracts: agency with moral hazard, risk neutrality, and limited liability. *RAND Journal of Economics*, 43(2), pp.215-234.

Roesler, A.K. and Szentes, B., 2017. Buyer-optimal learning and monopoly pricing. *American Economic Review*, 107(7), pp.2072-80.

Rogerson, W.P., 1985. The first-order approach to principal-agent problems. *Econometrica*, pp.1357-1367.