Endogenous Criteria for Success*

René Kirkegaard†

January 2022
Preliminary version for conference submission only

Abstract

Economic agents are motivated to undertake costly actions by the prospect of being rewarded for successes and punished for failures. But what determines what a success looks like? This paper endogenizes the criteria for success in an otherwise standard principal-agent model with a risk-neutral agent protected by limited liability. A contract now defines not only what the reward for success is, but also what constitutes a success in the first place. Under certain regularity assumptions, the more difficult it is to succeed, the fewer actions can be implemented. However, implementation costs decrease for those actions that remain implementable. Thus, the criteria for success can be used to manipulate implementation costs. If the criteria are of interest to the principal for only this reason, then the second-best action exceeds the first-best action. The opposite conclusion arises if the principal faces a budget constraint but all actions are implementable regardless of the criteria for success. When the principal's payoff depends directly and sufficiently strongly on the criteria of success themselves, then the second-best solution features either more stringent criteria for success or a lower action (or both) than the first-best solution.

JEL Classification Numbers: D82, D86

Keywords: Moral Hazard, Principal-Agent Models.

*I thank SSHRC for funding this research.
†Department of Economics and Finance, University of Guelph. Email: rkirkega@uoguelph.ca.

1 Introduction

The following new principal-agent model is proposed and studies. The agent's "performance" is a continuous variable whose distribution is determined by the agent's action. However, the principal is not able to perfectly observe performance. For example, consider a salesman (agent) who has been instructed to sell a product at some price p. The agent's efforts at persuading the customer will make the latter revise his willingness-to-pay for the product. In other words, the agent's "performance" is described by the customer's resulting willingness-to-pay. However, the customer's willingness-to-pay is inside his head and cannot be observed by outsiders. It can only be observed whether he decides to pay p or not. In other words, the agent and, more to the point, his employer (principal) know only whether the willingness-to-pay is above or below p. Thus, even though "performance" is continuous, the observable signal on which remuneration is based is binary. Moreover, the criterion for success is endogenously determined by p.

An equivalent situation arises when performance can be observed, but the principal is restricted in the reward structure. This often occurs when a pass/fail test is taken. The examiner may have a finer idea of the examinee's performance, but he cannot act on it. However, the criterion for success – the pass mark or the difficulty of the test – is often endogenous. For instance, a regulator may determine what it takes to pass a driver's license test, but leaves it to a certified middleman to determine if the applicant met those criteria.

Performance is one-dimensional in the model and the criterion for success is therefore essentially a performance threshold. If this is exogenous, the model reduces to a standard two-outcome model. However, when it is endogenous, it is entirely possible that the optimal threshold depends on the action that the principal wishes to induce. Note that the standard literature typically does not worry about where the probability of success comes from. One way of thinking about the current model is that a "microfoundation" of sorts is provided, linking the endogenous probability of success to an underlying performance technology. The model thus allows us to ask whether the criterion for success is more or less demanding as varying levels of effort is induced.

The threshold is fixed from the agent's perspective. Thus, his problem takes the same form as in the standard two-outcome model. Until recently, the literature has

imposed assumptions on the probability of success that ensures that the first-order approach (FOA) can be used. This is even less palatable in the current environment, where the probability of success is linked to the underlying performance technology. However, Kirkegaard (2017) offers a full solution to the two-outcome model without relying on the FOA (this is a special case of his model). Thus, the principal's problem can be solved for any given threshold and any underlying performance technology. A second step then requires searching for the best combination of threshold and action to induce.

To illustrate, consider once again the salesman and his employer. The first-best involves some effort level and a price that is set to maximize monopoly profits. However, under asymmetric information, two sources of distortions from the first best are possible. The employer may decide to distort the action or the price in order to lower implementation costs. Thus, the model is richer than the standard two-outcome model, which misses the dependence between the principal's choice variable (the price) and the probability of success.

The agent is assumed to be risk-neutral and protected by limited liability. This makes it possible to characterize implementation costs in a succinct and tractable way. Three versions of the problem are then analyzed.

The first version is in keeping with the existing literature. The FOA is assumed to be valid but the principal faces a budget constraint or wage cap. The second version focuses on the "implementability constraint" by studying environments where the FOA is not valid. This means that there are actions that cannot be implemented with just any threshold.

These two versions of the problem give contrasting conclusions. In other words, whether the binding constraint is a budget constraint or an implementability constraint matters. This is most obvious in cases where the principal is not directly interested in the criteria for success and only uses them to manipulate implementation costs. When the binding constraint is the budget constraint, the second-best action is lower than the first-best action. However, when the bidding constraint is the implementability constraint, the second-best action is *higher* than the first-best action. In both cases, the threshold is higher than the threshold that optimally implements the first-best action.

In the first two versions of the problem, either the budget constraint or the implementability constraint is binding. The reason is that the principal uses the criteria for

success with the sole purpose of manipulating implementation costs, thus driving her to select a contract that is on the boundary of the feasible set. The two versions give different conclusions because the feasible sets have dramatically different shapes. In the third version of the problem, the principal is assumed to intrinsically care about the criteria for success. In the salesman example, the principal is intrinsically interested in the price because this impacts her expected revenue. In this version of the model, the second-best solution need not be on the boundary of the feasible. However, it has more stringent criteria for success or a lower action (or both) than the first-best solution, even when allowing for both budget constraints and implementability constraints.

Two closely related papers are Bond and Gomes (2009) and Li and Yang (2020), although they are related for different technical and conceptual reasons.

Li and Yang (2020) consider the design of an optimal monitoring technology in which performance can be partitioned into an exogenously fixed number of categories. This is more general than the current paper, which only allows a partition into success and failure. However, Li and Yang (2020) concentrate on a situation in which there are only two effort levels. Among their extensions, they consider an environment with a finite number of effort levels, but they simply assume that the principal wishes to induce the highest possible action. Thus, they do not study what determines the optimal action, or how the first-best and second-best actions differ.

Bond and Gomes (2009) consider a setting in which the agent supply works on a number of independent tasks, each of which either succeeds or fails. The agent is compensated based on how many tasks succeed. Bond and Gomes (2009) show that the first-order approach is not valid in their setting. The problem is that the agent may want to deviate to the lowest possible effort profile (shirking on all tasks), and this incentive constraint therefore plays a critical role in their analysis. The first-order approach is generally not valid in the current paper either, and the problem is often, though not always, precisely that the agent is tempted to deviate to the lowest possible action. Thus, Bond and Gomes (2009) and the current paper are among the few that study the economic consequences of the failure of the first-order approach.

The contracting problem in Bond and Gomes (2009) results in two kinds of inefficiencies: A distortion in the total amount of effort, and a distortion in how this total amount is distributed among tasks. They assume that the first-best solution involves the agent working as hard as possible on each task. Thus, the only possible

distortion in total effort is downwards; the second-best effort must be no higher than the first-best. No such assumption is made in this paper, and it is in fact often the case that the second-best action is higher then the first-best action. As for the second distortion, the agent in Bond and Gomes (2009) concentrates too much of the total effort on a small number of tasks. In the model presented here, the second dimension of inefficiency comes from the fact that the principal distorts the criteria for success to manipulate the agent. In some settings, the criteria for success are themselves intrinsically important to the principal. In such cases, the principal often makes the criteria more stringent than she would ideally like, since this turns out to make it cheaper to induce effort.

In some ways, the problem in the current paper is simpler than in Li and Yang (2020) and in Bond and Gomes since only partitions into successes and failures are allowed and there is only one task. On the other hand, the implementability problem is tackled in more generality and distortions from the first-best are treated more carefully as it is not assumed that distortions in effort can only be downwards.

2 Incentive compatible contracts

The principal (she) employs a single agent (he). The agent takes some costly and unobservable action, a, belonging to some compact interval $[\underline{a}, \overline{a}]$. Given the agent's action, his performance is a random variable, X. Let $F(x|a)$ denote the corresponding distribution function, given a. For all $a \in (\underline{a}, \overline{a}]$, assume that there are no mass points, the support is the same interval for all $a \in (\underline{a}, \overline{a}]$, and the density $f(x|a) = F_x(x|a)$ is strictly positive on this interval, $[\underline{x}, \overline{x}]$, which may be bounded or unbounded.[1] At $a = \underline{a}$, the distribution either (i) satisfies these same assumption or, if $\underline{x}$ is finite, it (ii) is degenerate at $\underline{x}$. In the latter case, the agent's performance is guaranteed to be the worst possible if he takes the lowest possible action. This special case is included because it has some special implications that are useful in illustrating the workings of the model. The derivatives F_a, F_{aa}, f_a and their partial and cross partial derivatives are assumed to exist for all $a \in [\underline{a}, \overline{a}]$ in case (i) and for all $a \in (\underline{a}, \overline{a}]$ in case (ii). Throughout, it is assumed that $F_a(x|a) < 0$ for all $x \in (\underline{x}, \overline{x})$ and all $a \in [\underline{a}, \overline{a}]$ or $a \in (\underline{a}, \overline{a}]$ in case (i) and (ii), respectively. This assumption implies that actions are productive, in the sense that bad performances are less likely the harder the agent

[1] A subscript indicates the partial derivative with respect to the subscripted variable.

works.

Actions are normalized such that the agent's cost function is linear. Thus, the agent incurs a cost of a when he takes action a. Alternatively, think of the agent's action as a decision of what costs to incur. The agent is assumed to be risk neutral and protected by limited liability. This assumption makes it possible to succinctly characterize implementation costs, but it is not important for the more fundamental discussion of which contracts are incentive compatible.[2] The minimum wage implied by the limited liability constraint is normalized to zero. The agent's outside option is assumed to be so poor that the participation constraint never binds.

The principal either does not directly observe the agent's performance or his performance is not verifiable. Instead, what is observable and verifiable is whether the agent's performance exceeds a threshold x. Thus, $[\underline{x}, \overline{x}]$ is partitioned into two intervals, $[\underline{x}, x)$ and $[x, \overline{x}]$. The agent fails if his performance falls in the first interval and succeeds otherwise. Note that the threshold x in this way describes the criterion for success. The higher x is, the more stringent is the criterion.

Thus, there are two verifiable outcomes. The novelty is that the criterion for success is endogenized. That is, the principal controls what the threshold is. The stringency of the criterion affects the agent's incentives and therefore the implementation costs that the principal faces. The remainder of this section focuses on the incentive compatibility problem. Thus, no other restrictions are imposed on the contracting problem. However, the next section adds the possibility of a budget constraint.

To understand the agent's incentives, fix an interior threshold x and a bonus b that is paid out if the outcome is a success. Given the participation constraint is slack, it is optimal for the principal to pay nothing if the outcome is a failure. Then, the agent's expected utility from action a is

$$u(a|x, b) = b\,(1 - F(x|a)) - a,$$

since the probability of a success is $1 - F(x|a)$. Now assume that the principal is aiming to induce some specific interior action while holding fixed the threshold x. Then, b must be calibrated to ensure that the agent's first-order condition is satisfied

[2] As long as the agent has quasi-linear utility, the bonus from succeeding can be interpreted as being measured in "utils" rather than in monetary terms. The characterization of incentive compatible contracts carries through under this assumption.

at the intended action, or

$$-bF_a(x|a) - 1 = 0.$$

In other words, the bonus must equate

$$B(a,x) = \frac{-1}{F_a(x|a)}.$$

Hence, if the agent takes action a when offered the bonus $B(a,x)$, his expected wage is

$$W(a,x) = -\frac{1 - F(x|a)}{F_a(x|a)}$$

and his expected utility is

$$U(a,x) = W(a,x) - a.$$

It is useful to think of a contract as a triplet $(a, x, B(x,a))$, specifying a recommended action, a criterion for success, and a bonus if the outcome is a success. However, since $B(a,x)$ is uniquely nailed down by (a,x), the contract can summarized entirely by the pair (a,x). Here (a,x) can be read as: "the principal intends to induce action a by specifying threshold x and committing to the bonus $B(a,x)$." Now, the problem is that the first-order condition is necessary but not generally speaking sufficient for the agent's utility to attain a global maximum at the intended action a. In other words, there is more to the incentive compatibility problem. Thus, keep in mind that $W(a,x)$ and $U(a,x)$ are valid only as long as (a,x) is incentive compatible.

Fortunately, following Kirkegaard (2017), the fact that outcomes are binary makes it possible to completely characterize the incentive compatible contracts, or (a,x). To this end, fix $x \in (\underline{x}, \overline{x})$ and think of it as a parameter. Then, for a fixed bonus b, the curvature of the agent's expected utility depends only on the curvature of $1 - F(x|\cdot)$ with respect to a. Hence, the problem is locally concave in a if F is locally convex in a, or $F_{aa} \geq 0$. Now, starting from the function $F(x|\cdot)$, construct the convex hull (as a function of a), and denote this $F^C(x|\cdot)$. The convex hull is the largest convex function that lies on or below $F(x|\cdot)$. Thus, $F^C(x|a) \leq F(x|a)$ for all $a \in [\underline{a}, \overline{a}]$. For any x, let

$$A^c(x) = \{a \in [\underline{a}, \overline{a}] | F^C(x|a) = F(x|a)\}$$

denote the set of actions for which $F(x|a)$ coincides with $F^C(x|a)$. With some abuse of terminology, say that a is "on the convex hull" of $F(x|\cdot)$ if $a \in A^C(x)$. The end-points

of the domain are always on the convex hull, or $\underline{a}, \overline{a} \in A^C(x)$. Now, the contract (a, x) is incentive compatible if and only if it is true that there is no profitable deviation, or

$$U(a, x) \geq B(a, x)\,(1 - F(x|a')) - a'$$

for all $a' \in [\underline{a}, \overline{a}]$. Keeping in mind that $F_a(x|a) < 0$, a rearrangement yields

$$F(x|a) + (a' - a)\,F_a(x|a) \leq F(x|a').$$

Thus, the tangent line to $F(x|\cdot)$ through a must lie always below the function itself. This is the case if and only if $a \in A^C(x)$. In other words, (a, x) is incentive compatible if and only if $a \in A^C(x)$. Intuitively, $F^C(x|\cdot)$ is by construction a convex function. Thus, the agent's first-order condition is necessary and sufficient in an imaginary implementation problem where his technology is described by $F^C(x|\cdot)$ rather than $F(x|\cdot)$. Since $1 - F^C(x|\cdot) \geq 1 - F(x|\cdot)$, the agent's expected utility is at least as high in the imaginary problem as in the real problem. However, since $F^C(x|\cdot)$ and $F(x|\cdot)$ coincide on $A^C(x)$, it must also hold that if utility in the imaginary problem is maximized at some $a \in A^C(x)$ then it is maximized at the same action in the real problem. Examples are in Section 5.3 [to be completed].

Let

$$\mathcal{F} = \{(a, x) \in (\underline{a}, \overline{a}) \times (\underline{x}, \overline{x})\,|a \in A^C(x)\}$$

denote the set of *interior* (a, x) that are incentive compatible. This describes part of principal's feasible set, specifically those parts in which an interior action is induced. In comparison, to incentivize the actions at the corners, $\underline{a}$ and $\overline{a}$, the bonus is not pinned down by a first-order condition. The action $\underline{a}$ can be implemented with any threshold and a bonus of $B(\underline{a}, x) = 0$, since a zero bonus provides no incentive to work. Similarly, the action $\overline{a}$ can be implemented with any interior threshold $x \in (\underline{x}, \overline{x})$ by picking a bonus $B(\overline{a}, x)$ that is so large that the agent's utility is globally increasing in the action. Finally, thresholds of $\underline{x}$ or $\overline{x}$ can be used to induce only $\underline{a}$ since $F(\underline{x}|a) = 0$ and $F(\overline{x}|a) = 1$ are independent of a. These results are summarized in the next statement.

Lemma 1 *The set of implementable* (a, x) *is*

$$\mathcal{F} \cup \{(\underline{a}, x)|x \in [\underline{x}, \overline{x}]\} \cup \{(\overline{a}, x)|x \in (\underline{x}, \overline{x})\}.$$

Proof. In text. ■

The "implementability constraint" from now on refers to the condition that the principal must necessarily select a (a, x) contract that belongs to the set described in Lemma 1. All other contracts are not incentive compatible.

Recall that $B(a, x)$, $W(a, x)$, and $U(a, x)$ are uniquely nailed down by incentive compatibility on $\mathcal{F}$. This is not the case when $\underline{a}$ or $\overline{a}$ is induced. The interesting question in those cases is what the cheapest way to induce the action with threshold x is. For action $\underline{a}$, it is optimal and incentive compatible to offer a zero bonus, regardless of the threshold. Thus, let $B(\underline{a}, x) = W(\underline{a}, x) = 0$ and $U(\underline{a}, x) = -\underline{a}$. Note that implementation costs are generally discontinuous at $\underline{a}$ (an exception is considered in Section 5.2). Similarly, let $B(\overline{a}, x)$ denote the lowest bonus that can be used to induce action $\overline{a}$ with threshold $x \in (\underline{x}, \overline{x})$, and let $W(\overline{a}, x)$ and $U(\overline{a}, x)$ denote the resulting expected wage and expected utility, respectively. For now, it is sufficient to note that incentive compatibility may necessitate a higher bonus than what the first-order condition suggests and that a lower bonus would certainly cause the agent to deviate to a lower action. In other words, using the first-order condition gives *lower bounds* on $B(\overline{a}, x)$, $W(\overline{a}, x)$, and $U(\overline{a}, x)$.

Although the feasible set may have an unusual shape, it still possesses natural economic properties. First, holding fixed the threshold, a higher bonus must be offered to induce a marginally higher action on $\mathcal{F}$. The reason is that there is decreasing returns to scale in a on $A^C(x)$, implying that a bigger carrot has to be offered to entice higher effort. The decreasing returns to scale on $A^C(x)$ comes from the fact that the probability of succeeding, $1 - F^C(x|a)$, is concave in a in the imaginary problem, which coincides with the real problem on $A^C(x)$. Thus, when the agent is induced to work harder, he benefits not only from a higher bonus but also from a higher probability that he passes the fixed threshold. This double benefit increasing his expected wage and more than compensates for the fact that he also incurs higher effort costs.

Proposition 1 *The bonus, $B(a, x)$, as well as the agent's expected utility, $U(a, x)$, are weakly increasing in a on $\mathcal{F}$. The expected wage, $W(a, x)$, is strictly increasing in a on $\mathcal{F}$.*

Proof. See the Appendix. ■

So far, the problem has been analyzed for a fixed threshold. In particular, the set $A^C(x)$ traces out all the actions in $\mathcal{F}$, holding fixed x. Moving along the other dimension, let $X^C(a)$ denote the set of thresholds for which a can be implemented, $X^C(a) = \{x \in (\underline{x}, \overline{x}) \,|(a, x) \in \mathcal{F}\}$. Thus, $X^C(a)$ describes the set of thresholds that can be used to incentivize $a \in (\underline{a}, \overline{a})$.

3 Thresholds and implementation costs

The principal is assumed to be risk neutral. The cost $W(a, x)$ of implementing a feasible contract (a, x) depends on the threshold. Thus, even holding fixed the action, there is generally an incentive to manipulate the criterion for success in order to manipulate implementation costs.

However, the principal may also take a more direct interest in the threshold x. The expected benefit to the principal of (a, x) is $\pi(a, x)$. Her objective is therefore to maximize $\pi(a, x) - W(a, x)$ over the feasible set of contracts.

The benefit function is allowed to depend directly on the criterion for success. The leading example is $\pi(a, x) = (x - c)(1 - F(x|a))$. Here, the principal hires the agent to sell a product at price x. If successful, the principal incurs a cost c of supplying the product. Given an action a and a price x, the probability of a success is $1-F(x|a)$, and $\pi(a, x)$ thus describes expected profits. Note that $\pi(a, x)$ is generally non-monotonic in x in this example. The threshold x will be said to be "intrinsically important" to the principal whenever $\pi(a, x)$ depends on x. This encompasses situations in which it is harder or more costly for the principal to detect if performance exceeds some thresholds rather than others.

There are also contracting environments in which the principal does not care directly about the criterion for success. With some abuse of notation, the benefit function will be written more succinctly as $\pi(a)$ in those cases. The obvious example is $\pi(a) = \mathbb{E}[X|a]$. Here, the agent's performance can be interpreted as his productivity and the principal cares about his expected productivity. However, at the point in time at which the agent must be paid, it can only be verified whether the performance exceeded a pre-set threshold or not.

It is important to distinguish between the case where the threshold is intrinsically important to the principal and the case where it is not. In the latter case, the only role of the threshold is to manipulate implementation costs, whereas it serves a dual

purpose when it is intrinsically important.

Moreover, in the general case, social surplus is the difference between the benefits and the effort costs, or

$$S(a,x) = \pi(a,x) - a.$$

Thus, any first-best solution consists of a pair (a^{FB}, x^{FB}). The point here is that x is directly important for welfare. This in turn implies that any distortion in the threshold away from x^{FB} in the second-best problem is important for welfare reasons. This is not the case when the benefit function takes the form $\pi(a)$, in which case the first-best nails down an action, a^{FB}, while the threshold does not matter for social surplus at all. Stated differently, there is no unique first-best threshold in this case.

The principal's second-best problem is to maximize

$$V(a,x) = \pi(a,x) - W(a,x).$$

However, it is often more useful to think of the principal as the "residual claimant," since she receives what is left of social surplus after the risk neutral agent has receives his share, or

$$V(a,x) = S(a,x) - U(a,x).$$

Thus, it is important to understand the properties of $W(a,x)$ and $U(a,x)$ in order to solve the principal's second-best problem. For many of the following results, the monotone likelihood-ratio property (MLRP) is imposed. In fact, for expositional simplicity a strict version of the MLRP is used. Under the (strict) MLRP, the likelihood-ratio $\frac{f_a(x|a)}{f(x|a)}$ is (strictly) increasing in x. An equivalent definition is as follows.

DEFINITION (MLRP): The monotone likelihood-ratio property is satisfied if $f(x|a)$ is strictly log-supermodular in x and a, or

$$\frac{\partial^2 \ln f(x|a)}{\partial x \partial a} > 0.$$

Since log-supermodularity survives integration, it holds that the distribution function $F(x|a)$ as well as the survival function $1-F(x|a)$ are also strictly log-supermodular; see e.g. Athey (2002, Lemma 3). This observation is relevant because $W(a,x)$ can be written

$$W(a,x) = \left(\frac{\partial \ln\left(1 - F(x|a)\right)}{\partial a} \right)^{-1}.$$

This immediately implies that $W(a,x)$ is strictly decreasing in x. Note that $U(a,x)$ is strictly decreasing in x as well, since the threshold does not directly impact the cost of effort.

Proposition 2 *Assume MLRP is satisfied. Then, $W(a,x)$ and $U(a,x)$ are strictly decreasing in x on $\mathcal{F}$*

Proof. Differentiation yields

$$W_x(a,x) = -\left(\frac{\partial \ln(1-F(x|a))}{\partial a}\right)^{-2} \frac{\partial^2 \ln(1-F(x|a))}{\partial a \partial x} < 0,$$

which implies the result. ■

There is a potential trade-off when increasing the threshold. First, the higher the threshold is, the less likely it is that the bonus is paid out. On the other hand, it is possible that the bonus much be increased as well. The MLRP implies that the first effect dominates because the bonus increases relatively slowly.

When the criterion for success is not intrinsically important, the principal will aim to increase the threshold as much as possible in order to decrease implementation costs. This may give rise to an existence problem because a threshold of $\overline{x}$ is not incentive compatible – the agent never succeeds and will therefore pick action $\underline{a}$ in response.

There are at least three ways to deal with the existence problem. First, following the existing literature, the problem disappears under realistic restrictions on the set of permissible contracts. Specifically, in the current setting, a budget constraint or a wage cap makes it impossible to implement thresholds close to $\overline{x}$. Such contracting environments are considered in detail in Section 4. For future reference, let $\overline{b}$ denote the wage cap and let

$$\mathcal{F}_{\overline{b}} = \{(a,x) \in \mathcal{F} | B(a,x) \leq \overline{b}\}$$

denote the set of interior contracts that are both incentive compatible and satisfy the budget constraint. This describes the interior part of the feasible set in the second-best problem. This is potentially empty, so make the problem interesting it is assumed throughout that $\mathcal{F}_{\overline{b}} \neq \emptyset$. In other words, $\overline{b}$ is not too small. Note that $\mathcal{F}_{\infty} = \mathcal{F}$.

Second, under realistic assumptions on the technology, large threshold are not incentive compatible in the first place. That is, if $(a,x) \in \mathcal{F}$ then x is bounded away

from $\overline{x}$. This possibility is explored in Section 5. The dominant approach in the existing principal-agent literature downplays the incentive compatibility problem by assuming that the first-order approach is valid, or in other words that all interior (a, x) belong to $\mathcal{F}$. In contrast, the incentive compatibility problem is taken more seriously in Section 5, where it takes center stage. The conclusions of Sections 4 and Sections 5 are sometimes diametrically opposed.

Third, consider a setting in which the threshold is intrinsically important to the principal and the first-best threshold is in the interior. If she cares sufficiently much about the threshold, then distorting it too far away from the first-best is so undesirable that it does not compensate for the decrease in implementation costs that can be had from increasing it. To illustrate some basic arguments, a preliminary result in this vein is presented first. Assume that the first-best solution (a^{FB}, x^{FB}) is unique, interior, and that it can feasible be implemented in the second-best problem, or $(a^{FB}, x^{FB}) \in \mathcal{F}_{\overline{b}}$. Assume also that a solution to the second-best problem exists and that it is interior, i.e. in $\mathcal{F}_{\overline{b}}$. To understand how the first-best and second-best solutions compare, imagine that the principal in the second-best problem contemplates inducing (a^{FB}, x^{FB}). From this starting point, what are the consequences of changing the contract? First, any departure from (a^{FB}, x^{FB}) strictly lowers social surplus. Likewise, under the MLRP, any departure to another contract in $\mathcal{F}_{\overline{b}}$ that weakly increases a and/or weakly decreases x makes the agent at least weakly better off, by Propositions 1 and 2. Thus, such a (a, x) contract leaves strictly less surplus to the principal than she would get from inducing the first-best. In other words, the second-best cannot have both a larger a and a smaller x than the first-best.

Corollary 1 *Assume the MLRP holds. Assume that the first-best solution* (a^{FB}, x^{FB}) *is unique, interior, and in* $\mathcal{F}_{\overline{b}}$. *Assume that a second-best solution* (a^{SB}, x^{SB}) *exists and that it is in* $\mathcal{F}_{\overline{b}}$ *but different from the first-best.*[3] *Then, the second-best features either a strictly higher threshold than the first-best* ($x^{SB} > x^{FB}$), *a strictly lower action* ($a^{SB} < a^{FB}$), *or both.*

Proof. The proposition follows from the proof in the text that $x^{SB} \leq x^{FB}$ and $a^{SB} \geq a^{FB}$ cannot be jointly optimal. ■

Corollary 1 describes the possible ways in which the second-best is distorted away from the first-best when the agent is intrinsically interested in the criteria for success.

[3]The first-best and second-best may coincide in special cases, such as when $\pi(a, x)$ is discontinuous at (a^{FB}, x^{FB}).

More precise predictions depend on the properties of the benefit function. This is considered in Section 6. The next two sections focuses on contracting environments in which the principal is not intrinsically interested in the criteria for success.

4 Budget constraints

This section assumes that the principal faces a budget constraint. Thus, she can offer a bonus of at most $\overline{b}$, where $\overline{b}$ is bounded. To understand the implications of this restriction, note that the MLRP implies that the bonus $B(a,x)$ is u-shaped in x since

$$B_x(a,x) = \frac{f_a(x|a)}{F_a(x|a)^2}$$

is first negative and then positive as x increases.

One implication is that for a given action, a, the bonus $B(a,x)$ is minimized at the threshold, $x_0(a)$, for which the likelihood-ratio is zero, or $f_a(x_0(a)|a) = 0$. This is intuitive, because it means that the agent gets paid if and only if his performance is such that his action positively effects the probability of that performance, i.e. for all performances where $f_a \geq 0$. Thus, this threshold is where it is easiest to incentivize the agent. However, as Proposition 2 shows, expected wage costs decrease if the threshold exceeds $x_0(a)$ because the bonus, albeit higher, is then paid out less often.

More importantly, the u-shape implies that more extreme thresholds, whether they are high or low, require higher bonuses. The reason is that $F(x|a)$ is close to 0 or 1 when x is close to $\underline{x}$ or $\overline{x}$, respectively. Thus, it is harder for the agent to manipulate the chance of success. To entice him to work harder, the bonus must therefore be substantial. Recall that $F_a(x|a) < 0$ for $x \in (\underline{x}, \overline{x})$ but that $F_a(x|a) \to 0$ as $x \to \overline{x}$ or $x \to \underline{x}$. Hence, the bonus needs to grow without bound if sufficiently high or sufficiently low thresholds are used. In contrast, when x takes an intermediate value, $F(x|a)$ is easier to manipulate through a, and so a smaller bonus is required.

Consequently, the budget constraint implies that only thresholds in an intermediate range can be used to incentivize the agent. However, the set of thresholds that work depends on the action that the principal intends to induce. To focus on this issue, it is assumed, in line with most of the existing literature, that the first-order approach is valid, i.e. that all interior (a,x) belong to $\mathcal{F}$. Since this requires that $F(x|\cdot)$ coincide with its convex hull, it is necessary and sufficient that $F(x|a)$ is con-

vex in a. That is, Rogerson's (1985) Convexity of Distribution Function Condition (CDFC) is required.

DEFINITION (CDFC): The Convexity of Distribution Function Condition is satisfied if $F_{aa}(x|a) \geq 0$ for all x and all $a \in (\underline{a}, \overline{a}]$.

The CDFC is an oft-criticized condition. It is imposed here not because it is a desirable assumption but instead to focus squarely on budget constraints. The next section does the opposite, by ignoring budget constraints but relaxing the CDFC. The CDFC implies that the cheapest way to induce action $\overline{a}$ with threshold x is to use the bonus that satisfies the first-order condition. Thus, Propositions 1 and 2 also hold when $a = \overline{a}$.

To visualize the problem, start by sketching the iso-bonus curve where $B(a, x) = \overline{b}$ in (a, x) space. To start, fix the action a and assume that $B(a, x_0(a)) < \overline{b}$, meaning that there exists threshold that implements a while satisfying the budget constraint. Given that $B(a, x)$ is u-shaped in x, there is one threshold below $x_0(a)$ and one threshold above it for which $B(a, x) = \overline{b}$. Next, given the CDFC, it follows from Proposition 1 that $B(a, x)$ is weakly increasing in a. Thus, holding fixed the threshold, if $B(a, x)$ satisfies the budget constraint, then so do all smaller actions. In (a, x) space, the slope of the iso-bonus line where $B(a, x) = \overline{b}$ is $-\frac{B_a(x|a)}{B_x(x|a)} = -\frac{F_{aa}(x|a)}{f_a(x|a)}$. Thus, the iso-bonus curve slopes downward if $x > x_0(a)$ and upwards if $x > x_0(a)$. Figure 1 illustrates.

Note that the iso-bonus curve has two "prongs." Any (a, x) that is inside the area between the prongs satisfies the budget constraint and is therefore feasible. In contrast, Propositions 1 and 2 imply that iso-wage curves, $W(a, x) = c$, and indifference curves, $U(a, x) = c$, slope upwards on the feasible set $\mathcal{F}_{\overline{b}}$ when the MLRP is satisfied. Expected wages and expected utility are higher below the respective curves.

Assume that the benefit function $\pi(a)$ does not depend on the criteria for success. Let a^{FB} denote the first-best action and assume it is unique and interior. If $B(a^{FB}, x_0(a^{FB})) > \overline{b}$ then a^{FB} or any higher action is not implementable in the second-best problem. Then, the second-best action must be below the first-best action, or $a^{SB} < a^{FB}$.

Assume next that $B(a^{FB}, x_0(a^{FB})) \leq \overline{b}$. In the second-best problem and for any a, it is optimal for the principal to increase the threshold as much as is feasible. Hence, the optimal threshold is to be found on the upper prong or the downwards-sloping part of the iso-bonus line. Now imagine inducing a^{FB} with the highest possible threshold.

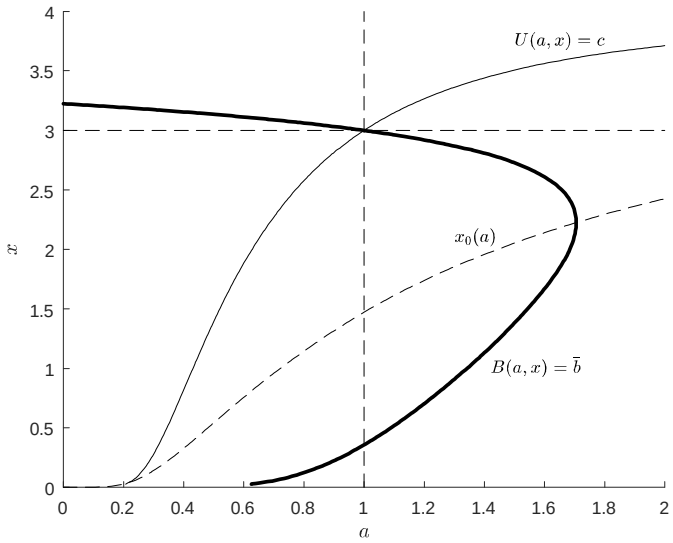

Figure 1: The feasible set and an indifference curve when $F(x|a) = \left(\frac{x}{4}\right)^a$, $x \in [0,4]$ and $a \in [0,2]$.

At this point, the weakly downward-sloping iso-bonus curve is intersected from below by a weakly increasing iso-wage curve. Thus, any feasible contract that involves a higher action is on or below the iso-wage curve, which means that it has both strictly lower social surplus and gives the agent weakly higher utility. This cannot be optimal for the agent. Thus, the second can be no higher than the first-best, or $a^{SB} \leq a^{FB}$.[4] This means that the threshold must be larger than the threshold that would optimally implement a^{FB}. Thus, the criteria for success are stringent. Indeed, since the second-best action is small and the threshold is large, there is a smaller probability that the agent is successful.

Proposition 3 *Assume that the MLRP and CDFC hold and that the principal is not intrinsically interested in the criteria for success. Assume the first-best action a^{FB} is unique and interior. Assume that the principal faces a budget constraint, $\overline{b} < \infty$. Then, any second-best action is no greater than the first-best action, $a^{SB} \leq a^{FB}$. If $\overline{b}$ allows a^{FB} to be implemented, then the second-best threshold is no smaller than the threshold that optimally implements a^{FB}.*

[4] If $\pi(a)$ is differentiable and $F_{aa}(x|a) > 0$ for $x \in (\underline{x}, \overline{x})$, then $a^{SB} < a^{FB}$. Intuitive, a small distortion away from a^{FB} has no first-order effect on social surplus but it has a first-order effect on the agent's expected utility.

Proof. In text. ■

5 Implementability constraints

The previous section imposed the CDFC, thus benefitting from the resulting conclusion that $\mathcal{F} = (\underline{a}, \overline{a}) \times (\underline{x}, \overline{x})$. In words, all interior (a, x) can be implemented. This section relaxes this assumption but instead ignores budget constraints. Although the CDFC is relaxed, some structure is still imposed on the problem. To set the stage, consider what is perhaps the simplest possible way to relax the CDFC, defined next.

DEFINITION (CAT): *Concavity at the top* is satisfied if, for any $a \in (\underline{a}, \overline{a})$, there exists some $x' \in (\underline{x}, \overline{x})$ such that $F_{aa}(x|a) < 0$ for all $x \in (x', \overline{x})$.

Concavity at the top (CAT) rules out the CDFC and implies that no interior action can be implemented with very high thresholds. Thus, this assumption solves the existence problem mentioned in Section 3.

Chade and Swinkels (2020) introduce a *no-upward-crossing* condition, which can be stated as the requirement that $F_{aa}(\cdot|a) - \tau F_a(\cdot|a)$ never crosses 0 from below on $(\underline{x}, \overline{x})$, for any $\tau \in \mathbb{R}$ and any $a \in A$. An equivalent statement is that $-F_a(x|a)$ is log-supermodular in a and x, or that $\frac{F_{aa}(\cdot|a)}{F_a(\cdot|a)}$ is increasing. Modifying Chade and Swinkels' (2020) terminology slightly, the abbreviation NUC_x will be used for no-upward-crossing with respect to x.

DEFINITION (NUC_x): The *no-upward-crossing* condition is satisfied if $-F_a(x|a)$ is log-supermodular in a and x.

It is an implication of NUC_x that $F_{aa}(\cdot|a)$ may be first positive and then negative as x increases. Thus, if NUC_x is satisfied but F is not convex in a for any x then CAT is automatically satisfied.

Chade and Swinkels (2020) provide sufficient conditions for NUC_x. Even though they provide counterexamples, they argue that NUC_x is a relatively weak condition. They mention the location families as a special example, such that $F(x|a)$ and $f(x|a)$ can be written as $Q(x-a)$ and $q(x-a)$, respectively. Here, it holds that $-F_a(x|a) = q(x-a) = f(x|a)$. Thus, in this case, the MLRP and NUC_x are the same condition. Incidentally, the condition is equivalent to requiring that q is log-concave.

CAT and NUC_x can be seen as a regularity condition that is consistent with Jewitt's (1988) condition. Jewitt (1988) assumes, among other things, that

$$\int_{\underline{x}}^{x} F(z|a)dz$$

is weakly convex in a for all x, or

$$\int_{\underline{x}}^{x} F_{aa}(z|a)dz \geq 0 \tag{1}$$

for all x. This is a natural assumption which implies that $\mathbb{E}[v(X)|a]$ is concave in a for any non-decreasing and concave Bernoulli utility functions $v(\cdot)$. Thus, this is a type of (expected) decreasing-returns-to-scale assumption that is weaker than the CDFC. From (1), $F_{aa}(x|a)$ must be non-negative when x is small but it can be negative when x is large. NUC_x says that it changes sign at most once, while CAT says that it must change sign at least once (thus precluding CDFC). Jewitt (1988) uses his assumptions to offer a justification of the first-order approach when performance is perfectly observable. However, this is not the case in the current setting, and the first-order approach is therefore not always valid. One way to understand the difference is to note that Jewitt's (1988) argument relies on additional assumptions that imply that optimal contracts give the agent utility that is concave in his performance. In the current setting, contracts are step-functions since the agent is compensated if and only if he passes the threshold. Such functions are not concave, and Jewitt's argument therefore does not apply.

The next subsection describes the shape of $\mathcal{F}$ under assumptions such as CAT, NUC_x, and another related assumption. The subsection following that studies the implications for the second-best problem.

5.1 The shape of the feasible set

Recall that NUC_x implies that $F_{aa}(\cdot|a)$ can be first-positive-then-negative as x increases. Thus, for a given action, F_{aa} is more likely to be negative the higher the threshold is. This suggests that the set of implementable actions shrinks as x increases. This intuition is correct, although the proof is slightly more involved as the convex hull of $F(x|\cdot)$ must be examined.

Proposition 4 *Assume NUC$_x$ holds. If $x', x \in (\underline{x}, \overline{x})$ and $x' > x$ then $A^C(x') \subseteq A^C(x)$. That is, fewer interior actions can be implemented the higher the threshold is.*

Proof. See the Appendix. ■

The proposition has some relevance to the literature that uses the FOA with a continuum of actions but two outcomes on the one hand, and the literature that assumes binary actions and two outcomes.

Corollary 2 *Assume NUC$_x$ holds. If $x', x \in (\underline{x}, \overline{x})$ and $x' > x$ then $A^C(x) = (\underline{a}, \overline{a})$ if $A^C(x') = (\underline{a}, \overline{a})$. In this case, the FOA is valid for any fixed threshold that is below x' (the first-order condition is sufficient for incentive compatibility). Likewise, $A^C(x') = \emptyset$ if $A^C(x) = \emptyset$. In this case, for any fixed threshold that is above x, only $\underline{a}$ and $\overline{a}$ can be implemented and the model reduces to a binary action model with two outcomes.*

Next, move along the other dimension. Thus, fix a target action and ask which thresholds can work to implement that particular action.

Proposition 5 *Assume NUC$_x$ and CAT are satisfied. Then, for any $a \in (\underline{a}, \overline{a})$, the set $X^C(a)$ is empty or it is an interval of the form $(\underline{x}, \overline{x}^C(a)]$, where $\overline{x}^C(a) < \overline{x}$. Thus, thresholds close to $\overline{x}$ cannot be used to implement a.*

Proof. See the Appendix. ■

For $a = \overline{a}$, let $\overline{x}^C(\overline{a})$ denote the highest threshold such that the bonus derived from the first-order condition is incentive compatible. Thresholds above $\overline{x}^C(\overline{a})$ can still be used to induce action $\overline{a}$, but the bonus must be made higher than what is suggested by the first-order condition.

NUC$_x$ implies that $A^C(x)$ shrinks when the threshold increases. In contrast, note that $X^C(a)$ may expand or contract as a increases. More regularity conditions are required in order to discipline $X^C(a)$. Hence, introducing a new definition, say that F satisfies *no-downward-crossing* with respect to a, abbreviated NDC$_a$, if $F_{aa}(x|\cdot)$ never crosses 0 from above on $(\underline{a}, \overline{a})$, for any $x \in [\underline{x}, \overline{x}]$. This allows $F_{aa}(x|\cdot)$ to be first-negative-and-then-positive as a increases. Note that this is a natural counterpart of NUC$_x$, which says something about the consequences of increasing x. Examples are in Section 5.3 [to be completed]

DEFINITION (NDC_a): *No-downward-crossing* with respect to a is satisfied if $F_{aa}(x|\cdot)$ never crosses 0 from above on $(\underline{a}, \overline{a})$, for any $x \in [\underline{x}, \overline{x}]$.

NDC_a implies that the set of feasible actions has a particularly simple structure.

Proposition 6 *Assume NDC_a holds. Then, for any $x \in (\underline{x}, \overline{x})$, the set of implementable actions takes either the form (i) $A^C(x) = \{\underline{a}, \overline{a}\}$, (ii) $A^C(x) = [\underline{a}, \overline{a}]$, or (iii) $A^C(x) = \{\underline{a}\} \cup [\underline{a}^C(x), \overline{a}]$, where $\underline{a}^C(x) \in (\underline{a}, \overline{a})$.*

Proof. See the Appendix. ■

Thus, if some interior action is implementable, then all higher actions are implementable as well. It is particularly important to note that if the action $\underline{a}^C(x)$ is induced, then the agent is exactly indifferent between the target action and the lowest action, $\underline{a}$. In case (i), define $\underline{a}^C(x) = \overline{a}$ and in case (ii) define $\underline{a}^C(x) = \underline{a}$. Then, in all three cases, the set $A^C(x)$ can be written in the form $\{\underline{a}\} \cup [\underline{a}^C(x), \overline{a}]$.

The next result combines the previous regularity assumptions. It can be thought of as describing the "nicest" shape of the feasible set that can be expected once the CDFC no longer holds.

Corollary 3 *Assume that CAT, NUC_x, and NDC_a all hold. Then, $\underline{a}^C(x)$ is weakly increasing in x on $(\underline{x}, \overline{x})$. Equivalently, $\overline{x}^C(a)$ is weakly increasing in a on $(\underline{a}, \overline{a})$.*

Proof. This follows from combining the conclusion that $A^C(x)$ shrinks when x increases with the conclusion that $A^C(x) = \{\underline{a}\} \cup [\underline{a}^C(x), \overline{a}]$. ■

5.2 The second-best solution

Consider the case in which the principal takes no direct interest in the threshold x. Thus, in the second-best problem, the threshold is manipulated with the sole purpose of lowering implementation costs. Given MLRP, wage costs are decreasing in x, and so the solution must be on the boundary of the feasible set.

To fix ideas, assume that there is a unique and interior first-best action, a^{FB}. Assume also that MLRP, CAT, NUC_x, NDC_a all hold. Assume that for any interior a, there exists an interior threshold $\overline{x}^C(a)$ such that a can be implemented if and only if the threshold is no larger than $\overline{x}^C(a)$.[5] Thus, $(a, \overline{x}^C(a))$ traces out the boundary

[5] CAT already ensures that $\overline{x}^C(a) < \overline{x}$. Thus, what is assumed in addition is that $X^C(a)$ is not empty. Hence, all a are implementable with some threshold. This rules out that $F(x|a)$ is globally concave in a for all x.

of the feasible set on $(\underline{a}, \overline{a}) \times (\underline{x}, \overline{x})$. Since wage costs are decreasing in x and π is independent of x, any interior solution to the second-best problem must be on the boundary, i.e. be of the form $(a, \overline{x}^C(a))$. Here, think of selecting the optimal action and the accompanying threshold $\overline{x}^C(a)$. From the perspective of social surplus, the threshold is irrelevant when π does not depend on x. Hence, this is good place to start in order to examine how the second-best action is distorted away from the first-best.

What is particularly important for the upcoming analysis is that NDC_a implies that the threshold $\overline{x}^C(a)$ has the following special property: If the action a is implemented with a threshold of exactly $\overline{x}^C(a)$ then the agent is exactly *indifferent* between a and the lowest possible action, $\underline{a}$. This is analytically convenient because the "anchor" $\underline{a}$ is independent of the action that is to be implemented, which in turn makes it easier to make inferences about the agent's utility along the boundary of the feasible set.

The best-case scenario for the principal is that $F(x|\underline{a}) = 1$ for all $x \in [\underline{x}, \overline{x}]$. An example is $F(x|a) = 1 - e^{-\frac{x}{\sqrt{a}}}$, with $x \geq 0$ and $\underline{a} = 0$. In words, the distribution is degenerate if the agent puts in the lowest possible action, in which case his performance is guaranteed to be the worst possible. Therefore, as long as x is in $(\underline{x}, \overline{x})$, there is no chance that the agent delivers a success, or earns the bonus, with action $\underline{a}$. This is advantageous to the principal, because such a deviation is less desirable and therefore easier to prevent. In particular, the aforementioned indifference condition is now

$$W(a, \overline{x}^C(a)) - a = -\underline{a}.$$

Thus, the agent is indifferent between all $(a, \overline{x}^C(a))$, $a \in (\underline{a}, \overline{a}]$. Stated differently, along the boundary of the feasible set, the agent appropriates a constant amount of rent and the rest goes to the principal. This is reminiscent of a standard principal-agent problem with risk neutral parties and a binding participation constraint. Thus, the first-best action solves the second-best problem. More formally, wage costs are

$$W(a, \overline{x}^C(a)) = a - \underline{a}.$$

Thus, if action a is implemented with threshold $\overline{x}^C(a)$, then the cost of implementation is $a - \underline{a}$. Similarly, $\underline{a}$ can be implemented with a zero bonus. At the other end of the support, for $\overline{a}$ there is no benefit to making the threshold exceed $\overline{x}^C(\overline{a})$ since the indifference condition must also hold at such thresholds (the binding IC constraint is

the no-jump constraint to $\underline{a}$). Thus, implementation costs are continuous in a.

In summary, the principal's payoff is

$$\pi(a) - a + \underline{a},$$

which by definition is no greater than $\pi(a^{FB}) - a^{FB} + \underline{a}$. In this form, the problem is easy to solve. Simply induce action a^{FB} by picking the threshold $x = \overline{x}^C(a^{FB})$. Thus, the first-best action is implemented, and it is implemented with the highest feasible threshold.

Proposition 7 *Assume MLRP, CAT, NUC$_x$, NDC$_a$ all hold. Assume that $F(x|\underline{a}) = 1$ for all $x \in [\underline{x}, \overline{x}]$. Assume that the principal's benefit function, $\pi(a)$, depends only on a and that there is a unique and interior first-best action, a^{FB}. Finally, assume that $\overline{x}^C(a)$ exists and is interior for all interior a. Then the second-best action coincides with the first-best action, $a^{SB} = a^{FB}$.*

Proof. The proof is in the text. ■

In this setting, there is no distortion of the optimal action. However, this is the case only because the threshold is endogenous and can be adjusted. If x is exogenous and fixed, the conclusion is much different. For instance, if x is fixed at a level that exceeds the solution to the second-best problem, then only $\underline{a}$ and actions above a^{FB} are feasible. Thus, the agent is either induced to take the lowest possible action or an action that exceeds the first-best. This example illustrated in a rather extreme way the benefit to being able to endogenize the criteria for success.

Next, remove the assumption that $F(x|\underline{a})$ is degenerate. Given MLRP, CAT, NUC$_x$, and NDC$_a$, it is still the case that the agent is indifferent between a and $\underline{a}$ when the threshold is $\overline{x}^C(a)$. However, a deviation to $\underline{a}$ now carries with it a strictly positive probability that the agent earns the bonus. Since the bonus depends on $(a, \overline{x}^C(a))$, the agent's utility therefore also depends on $(a, \overline{x}^C(a))$. In other words, the agent's utility is no longer constant along the boundary $(a, \overline{x}^C(a))$. This gives the principal an incentive to distort the action away from the first-best, since this allows her to manipulate the rent that has to be portioned off to the agent. It will now be proven that the action is distorted *upwards*, and that the threshold as a consequence must be "large."

The argument centers on comparing the agent's indifference curve to the boundary of the feasible set, summarized by the function $\overline{x}^C(a)$. As already noted, the indif-

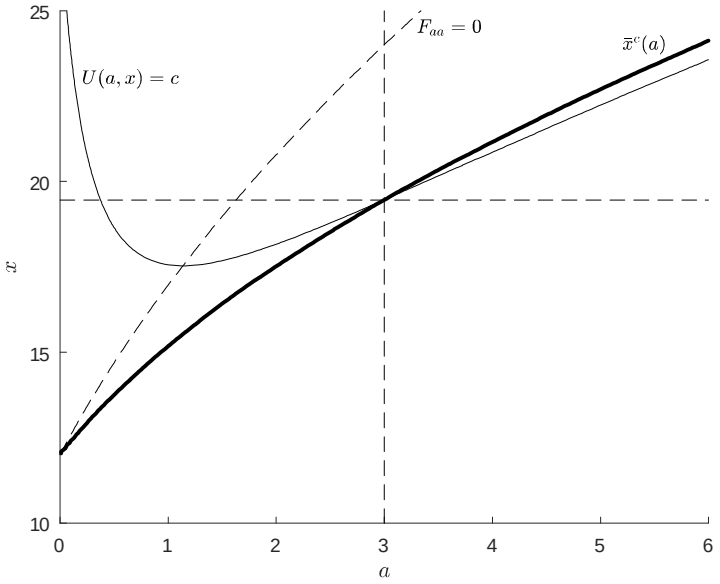

Figure 2: The feasible set and the agent's indifference curve.

ference curve must slope upwards on the feasible set, but so does $\overline{x}^C(a)$. However, it can be verified that at any point of intersection, the indifference curve is flatter than $\overline{x}^C(a)$. Thus, any indifference curve crosses $\overline{x}^C(a)$ at most once, and if so from above. In comparison, in the setting in the previous proposition, there is an indifference curve that coincides everywhere with $\overline{x}^C(a)$. Figure 2 illustrates this property, in the case where $F(x|a) = 1 - e^{-\frac{x}{4\sqrt{1+a}}}$, $x \geq 0$, $a \in [0, 6]$.

As in Figure 2, consider the indifference curve through the feasible point $(a^{FB}, \overline{x}^C(a^{FB}))$. Any feasible point below this curve is better for the agent and must have social surplus no higher than at $(a^{FB}, \overline{x}^C(a^{FB}))$ since social surplus is maximized whenever $a = a^{FB}$. Thus, all such points leave less surplus to the principal than does $(a^{FB}, \overline{x}^C(a^{FB}))$. The only feasible points that are potentially preferable feature actions and thresholds that are above a^{FB} and $\overline{x}^C(a^{FB}))$, respectively. Thus, if the second-best action is greater than $\underline{a}$ then it must be no smaller than the first-best action, $a^{SB} \geq a^{FB}$, and the optimal threshold must be larger than what is required to optimally implement the fist-best action, or $\overline{x}^C(a^{SB}) \geq \overline{x}^C(a^{FB})$. However, since implementation costs are discontinuous at $a = \underline{a}$, it cannot be ruled out that this is preferable to the principal. In conclusion, either $a^{SB} = \underline{a}$ or $a^{SB} \geq a^{FB}$.

Proposition 8 *Assume MLRP, CAT, NUC_x, NDC_a all hold. Assume that the principal's benefit function, $\pi(a)$, depends only on a and that there is a unique first-best action, a^{FB}, and that this is in the interior of A. Finally, assume that $\overline{x}^C(a)$ exists and is interior for all interior a. Then the second-best action is either $\underline{a}$ or it is no smaller than the first-best action, or $a^{SB} \geq a^{FB}$.*

Proof. See the Appendix. ■

The fact that the second-best action is *greater* than the first-best action is in very stark contrast to the conclusion that obtains from a standard model in which the threshold is fixed. Of course, if the fixed threshold is high enough then a^{FB} is not implementable; this occurs if $x > \overline{x}^C(a^{FB})$. Then, the second best is either $\underline{a}$ or, possibly, some action above a^{FB}. However, if a^{FB} is implementable, or $x \leq \overline{x}^C(a^{FB})$, then the second-best action must be *below* the first-best. The argument is by now familiar. Once again, the agent's expected utility is increasing in a. Hence, increasing a above a^{FB} decreases social surplus and increases the agent's surplus, thus leaving less surplus for the principal. Thus, in these cases, the standard model predicts that $a^{SB} \leq a^{FB}$. This is also the conclusion that was obtained when the threshold was made endogenous but the binding constraint was a budget constraint. Thus, even when the threshold is endogenous, it matter whether the budget constraint or the "implementability constraint" is binding. The reason is that the feasible set are shaped so differently; compare Figures 1 and 2.

5.3 Examples

To be completed.

6 Intrinsically important criteria for success

This section allows the benefit function to depend on the criteria for success and attempts to make more precise predictions than in Corollary 1. Thus, it starts by considering cases where neither the budget constraint nor the implementability constraint bind.

To be completed.

6.1 Interior first-best thresholds

The first-best threshold is interior in the leading example where $\pi(a,x) = (x-c)(1-F(x|a))$ and $c \in (\underline{x}, \overline{x})$. However, the analysis starts by examining the case where $\pi(a,x)$ is additively separable in a and x.

To be completed.

6.2 Monitoring and first-best thresholds at the corners

It is useful to clarify the relationship with Li and Yang (2020). They are interested in characterizing a monitoring technology that minimizes total costs, which consist of wage costs and the cost of the monitoring technology. The cost of the monitoring technology is increasing in the number of categories. In the current model, where the principal is restricted to at most two categories (success or failure), the cheapest thresholds are therefore $\underline{x}$ and $\overline{x}$ since in this case it collapses into only one category (the agent succeeds with probability one or fails with probability one, respectively). Li and Yang (2020) also assume that the naming of the categories do not matter for monitoring costs, so the cost is symmetric in the probability of success/failure. A monitoring cost function like

$$c(a,x) = -\left(F(x|a) - \frac{1}{2}\right)^2 \tag{2}$$

is consistent with their model. It is minimized at $\underline{x}$ and $\overline{x}$, and it is symmetric in the probability of success, $1-F(x|a)$, and the probability of failure, $F(x|a)$. However, Li and Yang (2020) devote particular attention to the special case in which the monitoring cost depends only on the number of categories. If the cost of additional categories is high enough, then two categories (success and failure) are optimal. The results in Sections 4 and 5 are then relevant. Unlike Li and Yang (2020) these results allow a comparison of first-best and second-best actions.

When monitoring costs are not constant, then thresholds $\underline{x}$ and $\overline{x}$ are the cheapest and there is an extra incentive to push the threshold towards one of the extreme values. If the principal's objective function is $\pi(a,x) = v(a) - c(a,x)$, then there is a first-best solution with $x^{FB} = \underline{x}$ and another with $x^{FB} = \overline{x}$. This is inconsistent with Corollary 1, where x^{FB} is assumed to be interior (see Section 6.1 for this case).

The first-best thresholds are not incentive compatible in the second-best problem.

Indeed, wage costs explode as the threshold approaches $\underline{x}$, and the budget constraint and/or the implementability constraint prevent the threshold from approaching $\overline{x}$. With a monitoring cost-function as in (2), there is an incentive to move away from a solution where $F(x|a) = \frac{1}{2}$. If the solution entails $F(x|a) > \frac{1}{2}$ then the budget or implementability constraint must bind, or else both wage costs and monitoring costs could be lowered by further increasing the threshold. If the solution entails $F(x|a) < \frac{1}{2}$, then it is because it balances higher wage costs against lower monitoring costs.

To be completed.

7 Conclusion

To be completed.

References

To be completed.

Bond, P. and A. Gomes, 2009, "Multitask principal–agent problems: Optimal contracts, fragility, and effort misallocation," *Journal of Economic Theory*, 144: 175–211.

Chade, H. and J.M. Swinkels, 2020, "The no-upward-crossing condition, comparative statics, and the moral-hazard problem," *Theoretical Economics*, 15: 445–476

Jewitt, I., 1988, "Justifying the First-Order Approach to Principal-Agent Problems," *Econometrica*, 56 (5): 1177-1190.

Li, A. and M. Yang, 2020, "Optimal incentive contract with endogenous monitoring technology," *Theoretical Economics*, 15: 1135-1173.

Kirkegaard, R., 2017, "Moral Hazard and the Spanning Condition without the First-Order Approach", *Games and Economic Behavior*, 102: 373-387.

Rogerson, W.P., 1985, "The First-Order Approach to Principal-Agent Problems," *Econometrica*, 53 (6): 1357-1367.

Appendix: Omitted proofs

Proof of Proposition 1. First, recall that any action in $\mathcal{F}$ is by definition interior. Next, fix a threshold x. The statement is trivial if $x \in \{\underline{x}, \overline{x}\}$ since only $\underline{a}$ can be implemented in that case. Thus, assume that $x \in (\underline{x}, \overline{x})$ and that $A^C(a)$ has at least two interior elements. Compare interior actions a and a' in $A^C(x)$, with $a' > a$. Since both actions are interior and on the convex hull of $F(x|a)$, it holds that $0 > F_a(x|a') = F_a^C(x|a') \geq F_a^C(x|a) = F_a(x|a)$. Thus, $B(a', x) \geq B(a, x) > 0$. Then, $1 - F(x|a') > 1 - F(x|a)$ implies that $W(a', x) > W(a, x)$. Finally, it follows from incentive compatibility and $B(a', x) \geq B(a, x)$ that

$$\begin{aligned} U(a', x) &= B(a', x)\,(1 - F(x|a')) - a' \\ &\geq B(a', x)\,(1 - F(x|a)) - a \\ &\geq B(a, x)\,(1 - F(x|a)) - a \\ &= U(a, x). \end{aligned}$$

■

Proof of Proposition 4. The result is trivial if $A^C(x) = (\underline{a}, \overline{a})$. Thus, assume $A^C(x) \neq (\underline{a}, \overline{a})$. Given an interior threshold, x, an action a^* in the interior of A is in A^C if and only if

$$F(x|a^*) + (a - a^*)\,F_a(x|a^*) \leq F(x|a)$$

or

$$F(x|a^*) - F(x|a) + (a - a^*)\,F_a(x|a^*) \leq 0$$

for all $a \in A$. Conversely, a^* is not in $A^C(x)$ if there exists any $a \in A$ for which the inequality is violated.

Since it is assumed that $A^C(x) \neq (\underline{a}, \overline{a})$, there must be some interior action that is not in $A^C(x)$. Select a^* to be this boundary point. Then, there exists some other action, a, for which the above inequality is violated, or

$$F(x|a^*) - F(x|a) + (a - a^*)\,F_a(x|a^*) > 0. \tag{3}$$

Differentiate the left hand side with respect to x to get

$$\begin{aligned}
\frac{\partial\left(F(x|a^*)-F(x|a)+(a-a^*)\,F_a(x|a^*)\right)}{\partial x} &= f(x|a^*)-f(x|a)+(a-a^*)\,f_a(x|a^*)\\
&= f(x|a^*)-f(x|a)-\frac{F(x|a^*)-F(x|a)}{F_a(x|a^*)}f_a(x|a^*)\\
&= \int_a^{a^*}\left(f_a(x|z)-\frac{F_a(x|z)}{F_a(x|a^*)}f_a(x|a^*)\right)dz\\
&= \int_a^{a^*}F_a(x|z)\left(\frac{f_a(x|z)}{F_a(x|z)}-\frac{f_a(x|a^*)}{F_a(x|a^*)}\right)dz\\
&= \int_a^{a^*}F_a(x|z)\left(\frac{f_a(x|z)}{F_a(x|z)}-\frac{f_a(x|a^*)}{F_a(x|a^*)}\right)dz.
\end{aligned}$$

Recall that $F_a(x|z)<0$. By NUC_x, $\frac{f_a(x|z)}{F_a(x|z)}$ is increasing in z. Thus, if $a^*>a$ then the term in the parenthesis is negative. Hence, the integral is positive. Similarly, if $a^*<a$, then

$$\frac{\partial\left(F(x|a^*)-F(x|a)+(a-a^*)\,F_a(x|a^*)\right)}{\partial x}=-\int_{a^*}^{a}F_a(x|z)\left(\frac{f_a(x|z)}{F_a(x|z)}-\frac{f_a(x|a^*)}{F_a(x|a^*)}\right)dz.$$

In this case, the term in the parenthesis is positive, and the right hand side is therefore positive as well. In other words, regardless of whether a is smaller or larger than a^*, an increase in x causes the term on the left in (3) to weakly increase. Thus, (3) is still satisfied. This means that a^* is still not implementable when x increases; it remains better to deviate to a. Thus, the set of implementable actions cannot grow as x increases. This proves the result. ■

Proof of Proposition 5. Fix an interior action a. Due to CAT, $F_{aa}(\cdot|a)$ must be negative when x is large enough. Such threshold are not incentive compatible and cannot implement a. Combined with NUC_x, $F_{aa}(\cdot|a)$ is therefore either always negative or it is first-positive-then-negative in x.

Given NUC_x, recall that if $x', x\in(\underline{x},\overline{x})$ and $x'>x$ then $A^C(x')\subseteq A^C(x)$. Thus, if $a\notin A^C(x)$ then $a\notin A^C(x')$. In words, if some threshold x cannot implement a then no higher threshold works either.

Combining the two observations implies that if $X^C(a)$ is not empty then it must take the form $(\underline{x},\overline{x}^C(a)]$, where $\overline{x}^C(a)<\overline{x}$. ■

Proof of Proposition 6. If $F_{aa}(x|\cdot)<0$ for all a then $A^C(x)=\{\underline{a},\overline{a}\}$ because

the extreme actions are the only actions on the convex hull of $F(x|\cdot)$ in this case. If $F_{aa}(x|\cdot) \geq 0$ for all a then $A^C(x) = [\underline{a}, \overline{a}]$. NDC_a permits only one additional possibility, namely that $F_{aa}(x|\cdot)$ is first negative and then positive as a increases. In this case, the set of actions on the convex hull of $F(x|\cdot)$ either consists only of $\{\underline{a}, \overline{a}\}$ or of $\underline{a}$ and an interval that extends to $\overline{a}$. In the latter case $A^C(x) = \{\underline{a}\} \cup [\underline{a}^C(x), \overline{a}]$, where $\underline{a}^C(x) \in (\underline{a}, \overline{a})$. ■

Proof of Proposition 8. The proof is outlined in the text, but it remains to prove that the indifference curve crosses $\overline{x}^C(a)$ at most once, and if so from above. It has already been shown that expected utility is globally decreasing in x. Similarly, expected utility is increasing in a on the feasible set, and indeed on the superset where $F_{aa}(x|a) > 0$. Thus, on this superset of the feasible set, the slope of the indifference curve is positive and equal to

$$\frac{dx}{da}_{U(a,x)=c} = \frac{F_{aa}(x|a)(1 - F(x|a))}{-f(x|a)F_a(x|a) - f_a(x|a)\,(1 - F(x|a))} > 0,$$

where the denominator is positive by the fact that the MLRP implies that the survival function $1 - F(x|a)$ is log-supermodular in (a, x).

Next, any point on the boundary of the feasible set, $(a, \overline{x}^C(a))$, is characterized by the condition that

$$F(x|a) - F(x|\underline{a}) + (\underline{a} - a)\,F_a(x|a) = 0.$$

This boundary has also been proven to have positive slope. The slope equals

$$\frac{d\overline{x}^C(a)}{da} = -\frac{(\underline{a} - a)\,F_{aa}(x|a)}{f(x|a) - f(x|\underline{a}) + (\underline{a} - a)\,f_a(x|a)} > 0.$$

Now compare the slopes at any point of intersection. Since it turns out to be easier

to compare the inverse functions, consider

$$\begin{aligned}
T(a,x) &= \frac{da}{dx}_{U(a,x)=c} - \frac{d\underline{a}^C(x)}{dx} \\
&= F_{aa}(x|a)\left[\left(-\frac{f(x|a)}{(1-F(x|a)}F_a(x|a) - f_a(x|a)\right) - \left(-\frac{f(x|a)-f(x|\underline{a})}{(\underline{a}-a)} - f_a(x|a)\right)\right] \\
&= F_{aa}(x|a)\left[-\frac{f(x|a)}{(1-F(x|a)}F_a(x|a) + \frac{f(x|a)-f(x|\underline{a})}{(\underline{a}-a)}\right] \\
&= F_{aa}(x|a)\left[-\frac{f(x|a)}{1-F(x|a)}F_a(x|a) + \frac{(f(x|a)-f(x|\underline{a}))\,F_a(x|a)}{F(x|\underline{a})-F(x|a)}\right] \text{ by definition of } (\underline{a}^C(x),x) \\
&= -F_a(x|a)F_{aa}(x|a)\left[\frac{f(x|a)}{1-F(x|a)} - \frac{f(x|a)-f(x|\underline{a})}{F(x|\underline{a})-F(x|a)}\right] \\
&= -\frac{F_a(x|a)F_{aa}(x|a)\times(f(x|a)\,(F(x|\underline{a})-F(x|a)) - (f(x|a)-f(x|\underline{a}))\,(1-F(x|a)))}{(1-F(x|a))\,(F(x|\underline{a})-F(x|a))} \\
&= -\frac{F_a(x|a)F_{aa}(x|a)\times(f(x|a)\,(F(x|\underline{a})-1) + f(x|\underline{a})\,(1-F(x|a)))}{(1-F(x|a))\,(F(x|\underline{a})-F(x|a))} \\
&= -\frac{F_a(x|a)F_{aa}(x|a)\,(1-F(x|\underline{a}))}{F(x|\underline{a})-F(x|a)}\left[\frac{f(x|\underline{a})}{1-F(x|\underline{a})} - \frac{f(x|a)}{1-F(x|a)}\right].
\end{aligned}$$

Since $F_a < 0, F_{aa} > 0$ and $F(x|\underline{a}) - F(x|a) > 0$, the term in front of the brackets in the last line is positive. By MLRP, $1-F(x|a)$ is log-supermodular in (a,x), implying that $\frac{f(x|a)}{1-F(x|a)}$ is decreasing in a. Therefore, the term in the brackets is positive too. Thus,

$$\frac{da}{dx}_{U(a,x)=c} - \frac{d\underline{a}^C(x)}{dx} \geq 0$$

or

$$\frac{d\overline{x}^C(a)}{da} - \frac{dx}{da}_{U(a,x)=c} \geq 0,$$

at any point of intersection. Thus, the indifference curve is flatter than $\overline{x}^C(a)$ in (a,x) space at any point of intersection.

The rest of the proof follows the argument in the text. ■