# Score-based mechanisms
## PRELIMINARY AND INCOMPLETE[*]

Eduardo PEREZ-RICHET[†]        Vasiliki SKRETA [‡]

14 March 2024

### Abstract

We propose a mechanism design framework that incorporates both soft information, which can be freely manipulated, and semi-hard information, which entails a cost for falsification. The framework captures various contexts such as school choice, public housing, organ transplant and manipulations of classification algorithms. We first provide a canonical class of mechanisms for these settings. The key idea is to treat the submission of hard information as an observable and payoff-relevant action and the contractible part of the mechanism as a mapping from submitted scores to a distribution over decisions (a score-based decision rule). Each type report triggers a distribution over score submission requests and a distribution over decision rules. We provide conditions under which score-based mechanisms are without loss of generality. In other words, situations under which the agent does not make any type reports and decides without a mediator what score to submit in a score-based decision rule. We proceed to characterize optimal approval mechanisms in the presence of manipulable hard information. In several leading settings optimal mechanisms are score-based (and thus do not rely on soft information) and involve costly screening. The solution methodology we employ is suitable both for concave cost functions and quadratic costs and is applicable to a wide range of contexts in economics and in computer science.

KEYWORDS: Mechanism Design, Fabrication, Ordeals, Manipulation, Cheating, Costly screening.

JEL CLASSIFICATION: C72; D82.

# Contents

# 1 Introduction

**Scoring Mechanisms** Financing decisions, admissions to selective higher education institutions and allocations of public housing units and of human organs are often performed via *score-based mechanisms.* These mechanisms rely on scores or priorities that measure and aggregate agents' characteristics into a one-dimensional score or metric. For instance, in consumer finance, credit scores assess an individual's creditworthiness and determine loan terms. In the financial sector, stress test scores evaluate the health of financial institutions. Education institutions use standardized test scores to evaluate student performance, while public school seats are allocated via priorities based on factors such as proximity to the school or a sibling in the school. In the medical field, human organ allocations are prioritized based on health status and treatments.

Oftentimes scores or priorities do not merely reflect agents' natural or true characteristics due to manipulations. As Frankel and Kartik (2019) point out, this leads to a distinction between an agent's natural score–obtained without interfering with the measuring technology–and the measured score, which may result from gaming, manipulation, or falsification. In school choice settings, there is ample evidence that some families submit fake addresses to achieve entry in desirable schools.[1] Doctors put their patients on escalated treatments in order to increase their priority on organ waiting lists (Bolton, 2018; McMichael, 2022). There is ample evidence of gaming of classification algorithms as discussed, for example, in Braverman and Garg (2020) and in the survey of Tang et al. (2023). Manipulations can be costly and the cost depends on hard information (agents' natural scores) and soft information (agents' abilities or tastes). Manipulations can impact the fairness and efficiency of these systems (see Hu et al., 2019).

We embark in Section 2 by proposing a unified framework that accommodates both soft and hard information. Traditionally in mechanism design, types represent soft information, allowing agents to lie freely. However, when introducing evidence, types are hard information and the standard assumption is that an agent either has a piece of evidence so the cost of lying is zero or does not have it in which case the cost of misrepresentation is infinite. By contrast, we allow for richer evidence structures captured by general falsification costs. We proceed to provide in Section 3 a canonical class of mechanisms for these settings. The key idea is to treat the submission of hard information as an observable and payoff-relevant action and the contractible part of the mechanism as a mapping from submitted scores to a distribution over decisions (a score-based decision rule). Each type report triggers a distribution over score submission requests and a distribution over score-based decision rules. This allows us to map this setting with costly misreporting to one captured by the generalized principal-agent setting in Myerson (1982) and obtain that truthful and obedient direct recommendation mechanisms (DRMs) are without loss of generality.

---

[1]For example, suggestive evidence indicates that parents fake addresses to gain admission to desirable public schools in Denmark (Bjerre-Nielsen et al., 2023).

Beyond the revelation principle, in our setting, we obtain two further simplifications that hold in general. First, DRMs can be decomposed into two mappings: each type report is mapped to a single *score-based decision rule* which, in turn, maps publicly submitted scores to a distribution over decisions and a *score recommendation rule* (a.k.a. falsification strategy) that maps a type report to random score recommendations. In other words, the mechanism does not need to randomize over score recommendation rules; randomization may be needed for score recommendations only. Second, because the agent submits scores publicly, obedience constraints are implied by voluntary ex-post participation constraints.

While the revelation principle in our setting identifies the language of inputs and outputs of the mechanism (type reports and score submission recommendations respectively) it does not specify that agents should be recommended to submit their natural scores. It is possible that optimal mechanisms request agents to submit falsified scores[2] Moreover, optimal mechanisms may involve random score submission recommendations. It is also possible that the same submitted score is assigned to a different distribution over final decisions depending on the type report that triggered it. For example, consider an agent whose natural score is their musical talent and soft information is their privately-known tastes for extracurricular activities. Depending on the designer's preferences the mapping from submitted music performance recordings to decisions (e.g., level of aid, major etc.) can vary with a student's type report and depend on student tastes.[3] Such mechanisms rely on a mediator and on reports from the agent.

In Section 4 we provide conditions under which conditioning the score-based decision rule on soft information is not needed and thus scoring-based mechanisms (that only condition on scores) are without loss of generality. In other words, we identify situations under which the agent does not make any type reports and decides without a mediator what score to submit in a score-based decision rule. A key condition is that the original mechanism does not rely on random score submission requests. This is de facto the case in settings in which the designer wants to ensure that agents submit their natural scores and restricts attention to falsification-proof mechanisms (see Perez-Richet and Skreta (2023) for example).

In Section 5 we characterize optimal mechanisms to screen a persuader in the presence of manipulable hard information. Put differently, we design optimal ordeal mechanisms or costly screening mechanisms. We do so using a methodology that is suitable both for concave cost functions and quadratic costs. In several leading settings the optimal mechanisms we characterize are implementable via score-based rules and do not even require commitment to the decision which can be taken by a third party, a decision maker that is a stand-in for consumers, firms and so forth. This new work provides a framework and methodologies applicable in a wide range of contexts in economics and in computer science.

---

[2]For an example, optimal tests in the presence of costly falsification leverage, what Perez-Richet and Skreta (2022) coin *productive falsification*, to improve the efficiency of decisions.

[3]See the example in Section 4.1 for an illustrative story.

## 1.1 Related literature

**Literature on mechanism design with evidence**  We contribute to the literature on mechanism design with evidence and in particular to that with moderate misreporting costs by providing a formulation that allows both for soft and hard information and by showing how this setting can be casted as a generalized principal-agent setting. Leveraging the revelation principle, we show how canonical mechanisms decompose into two mappings: a *score recommendation rule* and, a possibly, type-dependent *score-based decision rule*. The independent work by Schweighofer-Kodritsch and Strausz (2022) is in the same spirit but analyzes a setting with $0$ or $\infty$ costs and in which the implementable outcomes do not include the payoff relevant implications of presenting evidence.

*Infinite evidence costs:* In the classical formulation of mechanism design settings with evidence, an agent either has or does not have a piece of evidence ($0$ or infinity cost) and evidence is submitted as an input message in the mechanism see Green and Laffont (1986); Forges and Koessler (2005); Bull and Watson (2007); Deneckere and Severinov (2008); Ben-Porath and Lipman (2012). Green and Laffont (1986) are the first to note that the revelation principle fails in the sense that some social choice functions can only be implemented with partial evidence and provide conditions on the evidence structure, called nested range condition, under which the set of implementable social choice functions coincides with the set of truthfully implementable social choice functions. The subsequent papers Forges and Koessler (2005); Bull and Watson (2007); Deneckere and Severinov (2008) provide alternative conditions on the evidence structure available to agents (normality) such that presenting *maximal* evidence is without loss of generality recovering, in this weaker sense, the revelation principle. The reasons why truth-telling (in an appropriate sense) fails in settings with evidence, is simple: it not only matters which type(s) $t$ can mimic (call them $\mathcal{T}(t)$), but also which types in $\mathcal{T}(t)$ can mimic. Normality and the related conditions, guarantee that if $t$ can mimic $t'$, $t$ can also mimic any type $t'$ can mimic.

*Moderate evidence costs:* Bull (2008a) studies costly evidence production by two agents in a court setting and analyzes its effect of court outcomes when settlement and non-settlement are possible. Bull (2008b) allows for moderate linear evidence costs and shows that the sufficiency of the special three-stage mechanism of Bull and Watson (2007) holds also with moderate evidence cost. The mechanisms in Bull (2008b) differ from those in Bull and Watson (2007) in that they allow for a public signal from the external enforcer. Since transfers can be used to motivate the disclosure of evidence in the third stage, the second-stage signal can be public. However, transfers do not eliminate the need for the second stage because randomization by the external enforcer may be needed. Kartik and Tercieux (2012) study Nash implementation (as in Maskin, 1999) with evidence and their setting nests costly and hard evidence. By contrast, we show how mechanism design with evidence, regardless of the cost structure, can be cast with the original formulation of Myerson (1982).

**Literature on mechanism design with costly misreporting**  We also contribute to the literature on mechanism design with costly misreporting (Lacker and Weinberg (1989); Maggi and Rodriguez-Clare (1995); Crocker and Morgan (1998)) by showing that indeed the revelation principle in Myerson (1982) applies and by underlining the parts of the mechanism where randomization is needed. Lacker and Weinberg (1989) incorporate costly state falsification in a model of risk-sharing contracts and characterize optimal falsification-proof contracts, but also show they may be outperformed by contracts that induce falsification. Maggi and Rodriguez-Clare (1995), Crocker and Morgan (1998) derive mechanisms in settings with costly state falsification in single-agent settings with transfers. In Crocker and Morgan (1998) the contract specifies transfers and an action to be taken by the agent as a function of his type report $x$, denoted by $y(x)$. In some interpretations of their abstract model, the distance $|y - x|$ affects falsification costs. As is the case in Maggi and Rodriguez-Clare (1995), the optimal contract in Crocker and Morgan (1998) relies on distortions on $x$. Severinov and Tam (2019) focus on a mechanisms with transfers and provide conditions on reporting costs to ensure truth-telling is without loss. Deneckere and Severinov (2022) show that in environments with misrepresentation costs having agents send multiple signals, significantly expands the set of implementable outcomes and results to near efficiency when misrepresentation costs are small. Tan (2023) considers a price discrimination setting in which agents engage in costly behavior distortions to avoid being discriminated. By contrast, there are no transfers in Perez-Richet and Skreta (2022) who derive optimal tests in an agent-decision maker setting (a.k.a sender-receiver setting) in which the agent can falsify at a cost inputs into the test. Finally, there is also a computer science literature on mechanism design with reporting costs, most notably (Kephart and Conitzer, 2016) who provide conditions on reporting costs to ensure truth-telling is without loss.

**Literature on costly signaling and screening.**  We also relate to the vast literature stemming from Spence (1978)'s classic signaling model and the recent literature that studies costly gaming distortions in signaling settings stemming from the important contribution of Frankel and Kartik (2019). These works include Ball (2022) and Frankel and Kartik (2021). We differ in that we consider general mechanisms (rather than linear scoring rules). Our characterization of optimal approval mechanisms in settings without transfers relates to the literature on costly screening via various tools: notably money burning as studied, among others, in Hartline and Roughgarden (2008), Condorelli (2012), Chakravarty and Kaplan (2013). In those settings the utility burnt does not depend on the agent's type whereas in our setting falsification costs are type-dependent. Dworczak (2022) studies costly ordeals in a specialized setting with linear costs and allowing for deterministic mechanisms. In Dworczak (2022) the cost to achieve ordeal $y$ is linear in type. The allocation is an amount of money $x \in \mathbb{R}$ whereas we have an approval probability that must be in $[0, 1]$. This is not a crucial difference. In Dworczak (2022) and in most papers studying ordeals (costly screening mechanisms) the cost to achieve a given level of benefit is increasing in type. In our

setting, each type has a different ordeal level that costs zero (this is the ordeal level that corresponds to presenting the natural score). The optimal mechanisms we design simultaneously specify how each submitted score is mapped to an approval probability and a (random) score submission recommendation rule as a function of reported types. By contrast to these earlier works, the mechanisms we allow for can depend jointly on the ordeal and the type performing the given ordeal. This feature can be important for applications where the designer wants the allocation to depend jointly on the task and the type of agents because certain types have higher weight on designer's objective function.[4]

Our characterization of optimal approval mechanisms in settings without transfers also relates to Li and Qiu (2023) who study costly screening in a multi-agent setting without transfers and identify conditions under which contests are optimal and situations under which random mechanisms dominate contests. We differ in the cost structures we examine and the objective function.[5]

**Literature on persuasion mechanisms with evidence**    Glazer and Rubinstein (2004, 2006) study settings in which the sender seeks to persuade the receiver to 'accept' them regardless of the state of world whereas whether the receiver prefers to accept or to reject depends on the state of the world. The sender knows the state of the world and can present the receiver hard evidence about it. They derive optimal persuasion rules. These rules maximize the probability that the listener accepts the request if and only if it is justified. In Glazer and Rubinstein (2006) the authors show that neither commitment to the decision nor randomizations have any value. In Glazer and Rubinstein (2004) the sender can, in addition, send an arbitrary (cheap talk) message to the receiver and upon receiving the message the receiver can request hard evidence. Like in our example in Section 4.1, it is often beneficial for the receiver to randomise in the evidence she asks for from the sender. However, by contrast to our example in Section 4.1, in Glazer and Rubinstein (2004) it is not beneficial to randomise the final decision of accepting or rejecting once the evidence has been provided. Sher (2011) provides generalizations to the aforementioned results and identifies the key conditions on payoffs (namely, concavity) that renders commitment to have no value. Hart, Kremer, and Perry (2017) focus on truth-leaning equilibrium and identify the structure of evidence that guarantees that commitment cannot yield any advantage.[6]

---

[4] Akbarpour et al. (2023) analyze how to rank various costly screening tools. By contrast, we study given exogenously given falsification costs, how to optimally screen agents.

[5] As discussed in Li and Qiu (2023), mechanism design in the presence of costly manipulations, relates to works in computer science that study strategic classification. See, for instance, Hardt et al. (2016) who present an efficient classification algorithm that minimizes errors in the presence of gaming.

[6] Committing to a mechanism also has no value in the allocation setting without transfers considered by Ben-Porath et al. (2019).

## 2  Model

We consider an uninformed principal facing a privately-informed agent.[7] The agent's type, denoted as $t = (\theta, s) \in T$, consists of two components. First, $\theta \in \Theta$ represents the agent's preferences, abilities, and needs, which is soft information and can be misrepresented without cost. Second, $s \in S$ stands for the agent's natural score or evidence, such as a transcript or a certificate of disability, which is costly to falsify. There is a commonly-known prior over $T$, denoted by $F \in \Delta(T)$. The falsification cost function $c : A \times T \to \mathbb{R}_+$ defines the cost for a type $t = (\theta, s)$ to present a score $a \in A$, with $A \subseteq S$. We assume that presenting the true score is costless, i.e., $c(s, t) = 0; \forall t = (\theta, s)$. It is important to note that the cost of $a$ depends not only on the natural score $s$ but also on $\theta$, capturing aspects like gaming ability or discomfort from lying.

We denote by $X$ the set outcomes or decisions. An outcome $x$ can stand for quality-transfer pairs as in Mussa and Rosen (1978), levels of aid, bonuses, promotions and so forth. The agent's payoff is a function $u : X \times A \times T \to \mathbb{R}$, defined as

$$u(x, a, t) = v(x, t) - c(a, t).$$

Let $R$ denote the set of functions $\rho : T \to \Delta(A)$. In what follows, $\rho$ stands for the agent's falsification strategy. The designer's payoff is a function $u : X \times A \times T \times R \to \mathbb{R}$. Note that the designer's payoff does not only depend on the joint distribution of decisions and types (that is $X \times T$), rather it depends on the joint distribution over decisions, types, scores submitted as well as the agent's falsification strategy. It can be of the form

$$u_P(x, a, t, \rho) = v_P(x, t, a) - c_P(\rho, t),$$

capturing that the designer may internalize the agent's burnt utility caused from falsification. All the sets $T, A, S, \Theta$ are finite.

## 3  Canonical mechanisms

In this section we show how to cast our setting to a generalized principal-agent setting in Myerson (1982) and describe the canonical class of mechanisms.

To do so, we treat a score submission both as a type report and as a payoff-relevant action. An agent submits a type report which contains an informal report about their natural score, which is costless no matter the score they claim to have, and a formal submission of a score, which can be costly if they falsify their natural score. The formal score submission is payoff-relevant and analogous to an action in Myerson (1982). The revelation principle in Myerson (1982) establishes that any outcome, in other words, any joint distribution on $X \times A \times T$, arising at a BNE of any abstract mechanism, arises at a truthful and obedient equilibrium of a direct recommendation mechanism:

---

[7]We can straightforwardly extend the setting to accommodate multiple agents at the expense of more cumbersome notation.

A DRM maps type reports to a distribution over contractible outcomes (denoted by $D_0$ in Myerson, 1982) and action recommendations, one for each player.

The formal submission of a score $a \in A$ corresponds to choosing an action in Myerson (1982). The contractible outcomes in our setting, denoted by $D$, are functions from observable score submissions (actions) to distributions over decisions:

$$D = \{q : A \to \Delta(X)\}.$$

With our notation direct recommendation mechanisms are defined as follows:

$$\mu : \underbrace{T}_{\Theta \times S} \to \Delta(D \times A) \iff \mu : \underbrace{T}_{\Theta \times S} \to \Delta\left(\Delta(X)^A \times A\right).$$

Observe that each type report $t$ triggers a distribution over score requests ($a \in A$) and a distribution over score-based decision rules ($q \in D$). The mechanism then determines the ultimate joint distribution over types, decisions and submitted scores which matter because they are payoff relevant. Observe also that the mechanism may request falsified scores.

To summarize, in this formulation the agent reports a type $t = (\theta, s)$ and presents evidence $a$. A piece of evidence or *score* is submitted twice: first, informally via a costless message as part of the type report and second, as a formal submission which can be costly and amounts to choosing a payoff-relevant action in Myerson (1982). There are several real-world situations that resemble this procedure. In a job application setting, the costless message is to state a university degree on a CV and the formal submission is to show the certificate. In a disaster relief application, the costless message is the claim there was flood damage and the costly action is to produce evidence of the damage. Similarly, in a school choice setting the costless message amounts to stating the address in the form and the costly action amounts to producing evidence of this address (a utility bill, lease or ownership contract).

As part of the type report, the agent can freely lie about the score but the formal submission can be costly when the natural score is falsified.

**Proposition 1 (Revelation principle)** *(Myerson, 1982) Any BNE equilibrium feasible joint distribution on $T \times X \times S$ arising at a BNE of any abstract mechanism, arises at a truthfull and obedient equilibrium of a direct recommendation mechanism.*

*Sketch of proof* As a reminder, to obtain the result we start with some indirect mechanism $\pi : \mathcal{R} \to \Delta(D \times \mathcal{M})$ where $\mathcal{R}$ ($\mathcal{M}$) stand for abstract input (output) messages. The agent employs a reporting rule $\sigma : T \to \Delta(\mathcal{R})$ and chooses an action $a \in A$ as a function of message $m \in \mathcal{M}$ received using action rule $\delta : \mathcal{M} \to \Delta(A)$. The agent's strategy consists of the reporting and the action rule which together with the indirect mechanism $\pi$ determine the outcome which is a joint distribution over $X \times T \times A$. Alternatively, the mechanism $\pi$ and the agent's reporting and action rule can be thought

of as transition probabilities[8]

$$\sigma : T \to \mathcal{R}; \pi : \mathcal{R} \to D \times \mathcal{M}; \delta : \mathcal{M} \to A$$

and the corresponding DRM is a transition probability built from composing $\pi, \sigma, \delta$ as follows:

$$\mu = \sigma \circ \pi \circ \delta : T \to D \times A.$$

It is straightforward to argue that if $(\sigma, \delta)$ are part of a Bayes-Nash equilibrium given the indirect mechanism $\pi$, truth-telling and obedience are optimal for the agent given the DRM $\mu$. Mechanism $\mu$ by construction results to the same joint distribution over $X \times T \times A$ and to the same transition probability $\rho : T \to A$ as $\pi, \sigma, \delta$ do.

Thus the proof of Myerson (1982) straightforwardly extends to the proposed setting. The fact that the designer's payoff can depend on the $\rho : T \to \Delta(A)$ is immaterial for the argument.

**Incentive-compatible DRMs**   Incentive compatibility constraints split into truth-telling and obedience constraints:

$$\sum_{q \in D} \sum_{a \in A} \mu(q, a \mid t) \left[ \sum_{x \in X} q(x|a)v(x,t) - c(a,t) \right] \geq \sum_{q \in D} \sum_{a \in A} \mu(q, a \mid t') \left[ \sum_{x \in X} q(x|a)v(x,t) - c(a,t) \right] \quad \forall \, t, t' \in T$$

$$\text{(TT)}$$

$$\sum_{q \in D} \mu(q, a \mid t) \left[ \sum_{x \in X} q(x|a)v(x,t) - c(a,t) \right] \geq \sum_{q \in D} \mu(q, a \mid t) \left[ \sum_{x \in X} q(x|a')v(x,t) - c(a',t) \right] \forall \, t \in T; a \in \text{supp } \mu(\cdot|t), a' \in A.$$

$$\text{(OB)}$$

If the agent's participation in the mechanism is voluntary then the mechanism must also satisfy participation constraints:

$$\sum_{q \in D} \mu(q, a \mid t) \left[ \sum_{x \in X} q(x|a)v(x,t) - c(a,t) \right] \geq \underline{u}(t) \quad \forall \, t \in T, a \in \text{supp } \mu(\cdot|t) \quad \text{(PC)}$$

where $\underline{u}(t)$ denotes the agent's payoff from non-participation. For future reference we also let:

$$U(t, t', \mu) \equiv \sum_{q \in D} \sum_{a \in A} \mu(q, a \mid t') \left[ \sum_{x \in X} q(x|a)v(x,t) - c(a,t) \right]. \tag{1}$$

Earlier works on mechanism design settings with costly misrepresentation of types e.g. (Kephart and Conitzer, 2016; Severinov and Tam, 2019) treated the score submission analogous to a type report and identified conditions under which submitting the natural score is without loss of generality. With this more traditional view, the revelation principle (which is often used as a synonym with truth-telling) could fail. Indeed,

---

[8]When we refer to a mapping as transition probability from one set to another we remove the $\Delta$.

as discussed in the introduction, in settings without monetary transfers such as the ones considered in Perez-Richet and Skreta (2022) restricting attention to mechanisms that incentivize the agent to submit the natural score are with loss of optimality. This is true even for falsification cost functions that satisfy the triangular inequality considered in Kephart and Conitzer (2016) among others. Instead, here by treating the score submission as a payoff-relevant action we recover the generalized revelation principle by Myerson (1982). However, now, while the agent is reporting truthfully, the score submission request rule may ask the agent to falsify and it is not easy to identify a priori the optimal falsification targets for each type. We proceed to obtain further simplifications that we later leverage to solve for optimal mechanisms.

**Lemma 1 (Wlog deterministic mechanisms over score-based decision rules)** *Within our formulation, there is no need to randomize over score-based decision rules: each type report triggers a single $q \in D$.*

The intuition behind Lemma 1 is simple: Given that a score-based decision rule $q : A \to \Delta(X)$, specifies a randomization over $x$'s we do not need randomization over $q$'s: Conditional on a type report $t$ the mechanism specifies a unique score-based decision rule but possibly to random score submission requests.

**Mechanism decomposition**  In light of Lemma 1 mechanisms simplify from $\mu : T \to \Delta(D \times A)$ to $\mu : T \to D \times \Delta(A)$ and, more importantly, a mechanism $\mu$ decomposes into two mappings: A *score-based decision rule:*

$$q_\mu : A \times T \to \Delta(X)$$

defined from $\mu$ as follows $q(x|a, t) \equiv q(x|a)\mathbb{1}_{\mu(\cdot|t)=\delta_q}$, and a *score recommendation rule:*

$$\rho : T \to \Delta(A).$$

The mapping $\rho$ plays the role of the agent's falsification strategy and maps a type report to random score submission requests.

**Voluntary participation**  To formally accommodate the agent's participation decision we add additional actions in $A$. In what follows, we abuse notation and take $A$ to contain the scores that the agent submits and, in addition, the decisions to participate or not to participate.

In our setting, score submission is observable and hence, the mechanism can assign the null outcome or non-participation outcome (call it $\underline{x}$), if the agent inputs a score $a'$ in $q$ instead of the score $a$ requested by $\mu$. With this observation, the obedience constraints boil down to ex-post participation constraints. Here ex-post means conditional on a score submission recommendation rather than when we sum all recommendation on the support of $\rho$. We summarize this observation in the following lemma:

**Lemma 2 (Obedience implied by voluntary participation )** *Let $\underline{x}$ denote the non-participation outcome and let $\underline{u}(t)$ denote the corresponding payoff of type $t$. Within our formulation, obedience constraints reduce to participation constraints.*

Note that in the presence of a non-participation null outcome, the agent has to obey score submission requests that he is potentially not indifferent among.

**Simplified mechanisms**  From the above, it follows that it is without loss to focus on direct recommendation mechanisms with allocation rule $q_\mu : A \times T \to \Delta(X)$ and score recommendation rule $\rho : T \to \Delta(A)$ that satisfy *obedience:*

$$\mathbb{E}_{q_\mu(a,t)} v(x,t) - c(a,t) \geq 0, \qquad \forall a \in \operatorname{supp} \rho(t)$$

and *truth-telling*:

$$U(t) \equiv \mathbb{E}_{\rho(t)} \left( \mathbb{E}_{q_\mu(a,t)} v(x,t) - c(a,t) \right) \geq \mathbb{E}_{\rho(t')} \left\{ \mathbb{E}_{q_\mu(a,t')} v(x,t) - c(a,t) \right\}^+, \qquad \forall t, t'.$$

# 4   Score-based mechanisms

A *score-based mechanism* is an allocation rule based only on (final) score $a$, $q : A \to \Delta(X)$. A *falsification strategy* $\sigma : T \to \Delta(A)$ is incentive compatible if

$$\mathbb{E}_{q(a)} v(x,t) - c(a,t) \geq \mathbb{E}_{q(a')} v(x,t) - c(a',t), \quad \forall a \in \operatorname{supp} \sigma(t), \, a' \in A.$$

We proceed to provide sufficient conditions for a score-based mechanism to be without loss of generality.

**Assumption 1 (Separability assumptions)** *Suppose the agent's preferences can be written as*

$$u_A(x,a,t) = \beta(a,t) v(x) - c(a,t)$$

*with $\beta(a,t) > 0$ and the principal's preferences are:*

$$u_P(x,a,t) = w(a,t) r(x,a) + y(a,t) v(x) + \ell(a,t)$$

*with $w(a,t) > 0$.*

Note that the assumptions are automatically satisfied if $|X| = 2$ by normalization.

**Proposition 2 (Score-based principle)** *If preferences satisfy the separability assumptions and $(q_\mu, \rho)$ is an incentive compatible direct mechanism with a deterministic score recommendation rule, then there exists a score-based mechanism $q$ with an incentive compatible and falsification strategy $\sigma$ such that maintains the agent's payoff and weakly increases the principal's payoff.*

The result in Proposition 2 is analogous to the *taxation principle* whereby instead of submitting a report in a mechanism, the agent chooses an option from a menu. Here the agent only decides what score to submit in a fixed score-based decision rule, $\bar{q} : A \to \Delta(X)$. Instead in a mechanism, the allocation rule can depend on the agent's type as well, so $q_\mu : A \times T \to \Delta(X)$. In other words, in a mechanism two different types $t$ and $t'$ can be submitting the same score but face different distributions over decisions. When the mechanism relies on a deterministic score recommendation rule, incentive-compatibility implies that types $t$ and $t'$ are indifferent between each others distributions over decisions. The designer separability condition is necessary for the designer to always prefer one stochastic allocation for both types whenever two agent types submitting the same score are indifferent across two distributions over decisions. Principal separability is required because a stochastic allocation is multidimensional unlike a transfer (and if $|X| = 2$, no condition on the designer's payoffs is required). Agent separability plays an analogous role as the standard quasilinearity assumption.

**Implications** Suppose that the designer wants to restrict attention to falsification-proof mechanisms as is done in Perez-Richet and Skreta (2023)[9] then Proposition 2 implies that restricting attention to score-based mechanisms is without loss of generality.

**Definition 1 (Falsification proof mechanisms)** *An IC mechanism is falsification proof if it is a best response for all types $t = (\theta, s) \in T$ submit their true natural score, that is $a = s$.*

**Corollary 1** *In a multi-outcome generalized persuasion setting any IC falsification proof mechanism can be replicated by a scoring mechanism. Thus, the mapping from submitted scores to decisions cannot depend on soft dimensions of types.*

**Remark 1 (Comparison: tests versus score-based decision rules versus mechanisms)** *To understand the differences between general mechanisms versus score-based mechanisms considered in Perez-Richet and Skreta (2023) (which as stated in Corollary 1 are without loss in that setting) versus tests considered in Perez-Richet and Skreta (2022) we now compare them in our setting.*

*Test & falsification: Faced with a test $\tau : A \to \Delta(\mathcal{M})$ the agent chooses a falsification strategy that maps a natural score to a falsified score. The test converts the possibly falsified score to a signal $m \in \mathcal{M}$. There is a third party, a decision maker, who upon observing the signal $m$, makes a decision $X$.*

*Score-based mechanism & falsification: Now there is a score-based mechanism $q : A \to \Delta(X)$. Faced with such a mechanism the agent chooses a falsification strategy that maps a natural score to a falsified score. The mechanism maps a submitted score to a decision. There is commitment to the decision and no separate decision-maker.*

---

[9]Falsification-proof mechanisms do not burden agents and avoid negative externalities. In the words of Pathak and Sönmez (2008), falsification-proof mechanisms level the playing field.

*General mechanisms A general mechanism is*

$$q : T \times A \to \Delta(X) \text{ and } \rho : T \to \Delta(A)$$

*The agent submits a type report and the mechanism maps the report to a score-based decision rule and a score submission request. The revelation principle tells us that truth-telling and obedience are without loss. The difference between score based and general mechanisms is that in general q can vary with soft dimensions of type in contrast to score-based mechanisms.* ◇

We illustrate these differences in a simple example in what follows.

## 4.1 Illustrative example: college admission

The designer is a college facing a student with four equally likely types:

$$T = \{(F, s_L), (NF, s_L), (F, s_H), (NF, s_H)\}.$$

The first element of each type describes whether or not the student likes football (F or NF) whereas the second element is the natural score which can be low $(s_L)$ or high $(s_H)$. The decision is binary, so $X = \{0, 1\}$, where 0 stands for not admit while 1 stands for admit. The cost to falsify to $s_j \neq s_i$ is 1 for all types. The payoffs from each decision as a function of the agent's type are summarized in the table below where the first number is the agent's payoff whereas the second number is the designer's payoff:

|  | 1 | 0 |
|---|---|---|
| $t_1 = (F, s_L)$ | $1, 3$ | $0, 0$ |
| $t_2 = (NF, s_L)$ | $1, -1$ | $0, 0$ |
| $t_3 = (NF, s_H)$ | $1, 2$ | $0, 0$ |
| $t_4 = (F, s_H)$ | $1, 4$ | $0, 0$ |

Table 1: Student and college payoffs

 Scenario 1:  Suppose that the designer's payoff depends only on the admission decision. The designer's (here, the college's) first-best is to admit everyone except $t_2$. This can be achieved by faces the following test:

$$q(s_L) = 0, q(s_H) = 1$$

$t_2, t_3, t_4$ do not falsify; $t_1$ falsifies $s_L$ to $s_H$ and thus gets admitted but burns all utility because the cost of falsification is 1.  Scenario 2:  In this scenario, the designer's payoff depends on the decision and the incurred falsification costs. In particular, the loss function is quadratic incurred costs $c$ $L(c) = \frac{c^2}{6}$. Table 1 lists the payoffs corresponding to each decision. To get the total payoff for the designer we subtract the loss due to

falsification cost. Consider the following menu of score-based decision rules:

$$\text{pooling at 0: } q^0(s_L) = q^0(s_H) = 0$$
$$\text{pooling at 1: } q^1(s_L) = q^1(s_H) = 1$$
$$\text{separating: } q^S(s_L) = 0, q^S(s_H) = 1$$
$$\text{partially separating: } q^{PS}(s_L) = \frac{1}{4}, q^{PS}(s_H) = 1.$$

The optimal assignment to a score-based decision rule of each type is as follows:

$$\alpha(q^1|t_1) = 1$$
$$\alpha(q^{PS}|t_2) = 1$$
$$\alpha(q^S|t_3) = 1$$
$$\alpha(q^S|t_4) = 1.$$

Finally, the optimal score submission request rule is:

$$\rho(s_L|t_1) = \frac{1}{4}, \rho(s_H|t_1) = \frac{3}{4}$$
$$\rho(s_L|t_2) = 1, \rho(s_H|t_2) = 0$$
$$\rho(s_L|t_3) = 0, \rho(s_H|t_3) = 1$$
$$\rho(s_L|t_4) = 0, \rho(s_H|t_4) = 1.$$

Note that this mechanism satisfies truth-telling and obedience. Types $t_1, t_2$ get $\frac{1}{4}$ if they mimic each other and get zero if they mimic $t_3$ or $t_4$. Types $t_3, t_4$ get their maximum payoffs by reporting the truth. Whereas in Scenario 1 the optimum can be achieved by a test that entails no communication nor commitment on the decisions; in scenario 2 the optimum needs communication and commitment: it is a mechanism and not a test. In scenario 2, the designer's optimal mechanism *requests scores stochastically* and different type reports lead to *different* score-based rule. By contrast to the findings in Glazer and Rubinstein (2004), Sher (2011) and Hart et al. (2017) Ben-Porath et al. (2019) we have a pure persuasion setting with binary actions in which (i) commitment is valuable, (ii) communication is valuable, (iii) randomization in evidence requests is valuable and (iv) randomization in decisions is valuable. The difference with the earlier papers lies in that we consider general mechanisms that allow for randomizations and in that information in our setting is semi-hard and thus falsification is payoff-relevant. This example shows that scoring mechanisms (which, by contrast to tests encode commitment to the decision) can be dominated by a mechanism that receives type reports from the agent and outputs random score submission recommendations.

# 5 Optimally screening a persuader: binary outcomes

In this section we derive optimal mechanisms for the designer in a binary outcome setting in which $X = \{0, 1\}$ and agent types equal natural scores $T = [\underline{t}, \overline{t}]$, with $\underline{t} < 0 < \overline{t}$ so $t = s$ and there are no soft dimensions of the type. The distribution of types $F$ has full support and strictly positive density $f$. To fix ideas, we call decision 1 approval and decision 0 rejection. Regardless of type, the agent is a persuader who wants to be approved, so prefers $x = 1$ to 0, whereas the designer wants to approve only positive types. There are no transfers. This binary outcome setting captures many leading settings such as allocation of a good to an agent with unit demand without transfers (the agent either gets a unit or not); acceptance decisions, approvals, promotions and many other settings. As discussed in the introduction, analogous settings have been analyzed in Glazer and Rubinstein (2004), Sher (2011) and Perez-Richet and Skreta (2023). Li and Qiu (2023) study a richer allocation problem with many goods and agents under linear signaling costs.[10]

The agent, regardless of type, gets a payoff of 1 if approved and 0 otherwise, that is for all $t \in T$:

$$u(x, a, t) = v(x, t) - c(a, t) = \begin{cases} 1 - c(a, t) \text{ if } x = 1 \\ -c(a, t) \text{ if } x = 0 \end{cases}.$$

The designer's outside option from rejecting the agent is 0. The designer's payoff from accepting an agent of type $t$ is equal to $t$, therefore the first-best is to accept positive types and to reject negative ones. There are no resource constraints. The cost function is scaled by $\gamma > 0$ which stands for the agent's gaming ability assumed to be known, so $c : T \times A \to \mathbb{R}$ and $\frac{1}{\gamma}c(a, t)$ denotes the cost to type $t$ of submitting score $a$. Types are distributed according to a commonly-known full support distribution $F$. We focus on the case that $\mathbb{E}_F[t] < 0$. We assume that no falsification is costless so when $a = t$, $\frac{1}{\gamma}c(a, t) = 0$ (which is just a normalization) and that for all $a \geq t$ the cost is decreasing in $t$.

For this binary outcome setting, the allocation rule simplifies to an approval probability as a function of a type report $t$ and a submitted score $a$:

$$q(a, t) \in [0, 1].$$

The mechanism, therefore, consists of an allocation rule that depends both on the agent's type $t$ and submitted score $a$, $q : A \times T \to [0, 1]$ and score recommendation rule $\rho : T \to \Delta(A)$.

The interim approval probability is the expectation over all recommended scores:

$$Q(t) = \int_A q(a, t) d\rho(a, t). \tag{2}$$

---

[10]Under certain conditions, their findings relating to whether or not contests are dominated by mechanisms that involve randomness go through under convex costs as they explore in their appendix.

The agent's payoff simplifies to $U(t) = Q(t) - \int_{a \in A} \frac{1}{\gamma} c(a,t) d\rho(a|t)$ and the ex-post participation constraints (which, as mentioned earlier, imply obedience) and truth-telling constraints write:

$$q(a,t) - \frac{1}{\gamma} c(a,t) \geq 0 \quad \forall a \in \operatorname{supp} \rho(\cdot|t) \tag{PC}$$

$$U(t) \geq \int_{a \in A} \left[ q(a,t') - \frac{1}{\gamma} c(a,t) \right] d\rho(a|t') = Q(t') - \int_{a \in A} \frac{1}{\gamma} c(a,t) d\rho(a|t') \quad \forall t, t' \in T. \tag{TT}$$

**Designer's objective** The designer seeks the mechanism that solves

$$\max_{q, \rho} \int_{\underline{t}}^{\overline{t}} Q(t) t f(t) dt$$

subject to TT, PC, probability constraints.

In what follows, when a score recommendation rule $\rho$ is deterministic for each $t$ (a Dirac on some $a$) we denote it simply as $a^* : T \to A$.

**First best** The first-best for the designer is:

$$Q^{\mathrm{FB}}(t) = \begin{cases} 0 \text{ for } t < 0 \\ 1 \text{ for } t \geq 0. \end{cases} \tag{3}$$

If $1 - \frac{1}{\gamma} c(\overline{t}, 0) < 0$, so $\gamma < c(\overline{t}, 0)$ we can achieve the first best by setting

$$a^*(t) = \begin{cases} t \text{ for } t \in [t^\star, \overline{t}] \\ t^\star \text{ for } t \in [0, t^\star) \\ t \text{ for } t < 0 \end{cases} \qquad q(a,t) = \begin{cases} 1 \text{ for } a \geq t^\star \\ 0 \text{ for } a \neq t^\star \end{cases}$$

where $t^\star$ satisfies $1 - \frac{1}{\gamma} c(t^\star, 0) = 0$. In words, $t^\star$ is the highest score type 0 is willing to falsify to in order to get approved with probability 1. All positive types get approved with probability 1. Positive types up to $t^\star$ falsify to $t^\star$ while all positive types above $t^\star$ do not falsify. All negatives get approved with probability 0 and do not falsify.[11]

In what follows, we solve for the optimal mechanism when $1 - \frac{1}{\gamma} c(\overline{t}, 0) \geq 0$ so

$$\gamma > c(\overline{t}, 0) \tag{4}$$

holds and therefore falsification costs are low enough to make the first-best impossible.

**Lemma 3** *Without loss of optimality we can restrict attention to $q : A \times T \to [0,1]$ increasing in $a$ for all $t$.*

_____

[11]The definition of $t^\star$ and the fact that $c$ is decreasing in $t$ imply together that $1 - \frac{1}{\gamma} c(t^\star, t) < 0$ for $t < 0$.

## 5.1 Concave costs

Suppose that for each $a \in A$, with $a \geq t$, $c(\cdot, t)$ is decreasing and concave in $t$. Recall that in this section, $t \in T = [\underline{t}, \overline{t}] \subset \mathbb{R}$.

**Implications of incentive compatibility**  The agent's payoff from truth-telling writes:

$$U(t) = \max_{t' \in T} \left( Q(t') - \int_{a \in A} \frac{1}{\gamma} c(a, t) d\rho(a|t') \right)$$

and when $c$ is concave it is convex as it is the maximum of convex functions. Let

$$C(t) \equiv - \int_{a \in A} \frac{\partial \frac{1}{\gamma} c(a, t)}{\partial t} d\rho(a|t). \tag{5}$$

Note that if $Q$ can take any value in $\mathbb{R}$ then it is analogous to a transfer making our problem similar to a mechanism design problem with quasilinear payoffs. Lemma 4 that follows is standard.

**Lemma 4** *A mechanism $q, \rho$ satisfies truth-telling* $\iff$

1. *$C(t)$ is increasing*

2. *$C(t)$ belongs to the subgradient of $U$*

3. *$U(t) = U(\underline{t}) + \int_{\underline{t}}^{t} C(z) dz = U(\overline{t}) - \int_{t}^{\overline{t}} C(z) dz$.*

In our setting, however, $Q$ can only take values in $[0, 1]$. We proceed to build and solve a relaxed problem and in the process ensure that the $Q$ is in the correct range namely in $[0, 1]$. Combining (2) and the equality in item 3 above we can express $Q$ as follows:

$$Q(t) = U(t) + \int_{a \in A} \frac{1}{\gamma} c(a, t) d\rho(a|t) = U(\underline{t}) + \int_{\underline{t}}^{t} C(z) dz + \int_{a \in A} \frac{1}{\gamma} c(a, t) d\rho(a|t)$$

$$= U(\underline{t}) - \int_{\underline{t}}^{t} \left[ \int_{a \in A} \frac{\partial \frac{1}{\gamma} c(a, z)}{\partial z} d\rho(a|z) \right] dz + \int_{a \in A} \frac{1}{\gamma} c(a, t) d\rho(a|t). \tag{6}$$

Recall that in the case we are analyzing $\mathbb{E}_F[t] < 0$. Let $t_0$ be such that $\int_{t_0}^{\overline{t}} z dF(z) = 0$, so $t_0$ is the type above which the conditional expectation of the agent's type is equal to 0. The designer wants to minimize the approval probability for negative types. Because $Q(t)$ is increasing in $U(t)$ and in falsification costs, at an optimum there is a boundary type $t_* \geq t_0$[12] such that (i) $U(t) = 0$ for all $t \in [\underline{t}, t_*]$ (ii) $\rho(a|t) = \delta_t$ for all $t \in [\underline{t}, t_*]$; no falsification for these low types sets falsification cost to zero and ensures $Q(t) = 0$. With these observations (6) writes:

---

[12]We explain why the boundary type must be above $t_0$ below.

$$Q(t) = -\int_{t_*}^{t} \left[ \int_{a \in A} \frac{\partial \frac{1}{\gamma} c(a,z)}{\partial z} d\rho(a|z) \right] dz + \int_{a \in A} \frac{1}{\gamma} c(a,t) d\rho(a|t). \tag{7}$$

We employ (7) and standard arguments to rewrite the designer's objective of the reduced problem as follows:

$$\int_{t_*}^{\bar{t}} Q(t) t f(t) dt = \int_{t_*}^{\bar{t}} \int_{a \in A} \left( \frac{1}{\gamma} c(a,t) t - \frac{\partial \frac{1}{\gamma} c(a,t)}{\partial t} \mathbb{E}_F[z | z \geq t] \frac{(1 - F(t))}{f(t)} \right) d\rho(a|t) f(t) dt$$

and the principal's problem becomes:

$$\max_{q,\rho} \int_{t_*}^{\bar{t}} \int_{a \in A} \left( \frac{1}{\gamma} c(a,t) t - \frac{\partial \frac{1}{\gamma} c(a,t)}{\partial t} \mathbb{E}_F[z | z \geq t] \frac{(1 - F(t))}{f(t)} \right) d\rho(a|t) f(t) dt$$

subject to $C$ increasing and $Q \in [0,1]$

where $C$ is defined in (5).

**Relaxed problem** As usual, we solve the relaxed problem ignoring the monotonicity constraint on $C$ which is required for truth-telling. The designer's objective is linear in $\rho$'s. Moreover, in Lemma 3 we have established that the assignment probability $q(a,t)$ is increasing in $a$. Then, without loss of generality we can restrict attention to score submission recommendations weakly above the natural score $t$. The pointwise optimal $\rho$ is by construction deterministic (a Dirac on some $a$) and we denote it simply as $a^* : T \to A$. This optimal action recommendation solves for each $t \in T$ the following:

$$\max_{a \in A} \frac{1}{\gamma} c(a,t) t + \frac{\partial \frac{1}{\gamma} c(a,t)}{\partial t} \mathbb{E}[x | x \geq t] \frac{(1 - F(t))}{f(t)}.$$

**Monotonicity** Per Lemma 4 truth-telling requires that $C(t)$ is increasing in $t$. Using the pointwise optimal recommendation strategy which is deterministic, $C(t)$ becomes:

$$C(t) = -\frac{\partial \frac{1}{\gamma} c(a^*(t), t)}{\partial t}. \tag{8}$$

Note that $C$ is increasing so long as $-\frac{\partial \frac{1}{\gamma} c(a^*(t),t)}{\partial t}$ is increasing in $t$.

In what follows, we solve for the pointwise optimal $\rho$ and derive the corresponding optimal $q$ for two classes of falsification cost functions conditions (i) linear in distance falsification costs[13] and (ii) quadraric costs.[14] Linear costs are trivially concave. Quadratic costs are convex but the solution approach applies to this case as well as we explain below. In the linear cost case, the solution to the relaxed program is feasible

---

[13] Among others, Li and Qiu (2023) assume linear costs.
[14] Among others, Frankel and Kartik (2019, 2021) analyze quadratic costs.

for all distributions of scores. For the case of quadratic costs, the solution is feasible for scores distributions $F$ satisfying the monotone hazard rate property.

### 5.1.1 Linear cost

Suppose $\frac{1}{\gamma}c(a,t) = \frac{1}{\gamma}|a-t|$. As anticipated above, we solve for the optimal mechanism for the interesting range of $\gamma$ in which the first-best is not feasible. This is the range of gaming abilities such that (4) is satisfied, which for the linear costs becomes $1 - \frac{1}{\gamma}\bar{t} \geq 0$ or $\gamma > \bar{t}$.

For $a \geq t$, the cost is $\frac{1}{\gamma}(a-t)$ and its derivative is $\frac{\partial \frac{1}{\gamma}c(a,t)}{\partial t} = -\frac{1}{\gamma}$ and

$$C(t) = -\int_{a \in A} \frac{\partial \frac{1}{\gamma}c(a,t)}{\partial t} d\rho(a|t) = \frac{1}{\gamma}.$$

The principal's program for this cost function becomes:

$$\max_{q,\rho} \int_{t_*}^{\bar{t}} \int_{a \in A} \left( \frac{1}{\gamma}(a-t)t + \frac{1}{\gamma}\mathbb{E}[x|x \geq t]\frac{(1-F(t))}{f(t)} \right) d\rho(a|t)f(t)dt$$

subject to $C$ increasing and $Q \in [0,1]$

and the pointwise optimum solves:

$$\max_{a \in [t,\bar{t}]} \frac{1}{\gamma}(a-t)t + \frac{1}{\gamma}\mathbb{E}[x|x \geq t]\frac{(1-F(t))}{f(t)}.$$

The expression is linear in $a$ and the optimum is a corner solution

$$a^*(t) = \begin{cases} \bar{t} \text{ for } t \geq 0 \\ t \text{ for } t < 0 \end{cases} \quad \text{and} \quad C(t) = \begin{cases} \frac{1}{\gamma} \text{ for } t \geq 0 \\ 0 \text{ for } t < 0 \end{cases}$$

which is increasing. Thus, the monotonicity constraint is satisfied for all type distributions $F$.

For $\bar{t}$, $c(a^*(\bar{t})|\bar{t}) = c(\bar{t}|\bar{t}) = 0$, so $U(\bar{t}) = Q(\bar{t})$. We let $Q(\bar{t}) \equiv p^*(\gamma)$ and we proceed to identify its value below. The assignment probability for $t \geq 0$ is

$$Q(t) = U(t) + c(a^*(t)|t)$$

$$= U(\bar{t}) - \int_t^{\bar{t}} C(x)dx + c(a^*(t)|t)$$

$$= U(\bar{t}) - \int_t^{\bar{t}} \frac{1}{\gamma}dx + c(a^*(t)|t)$$

$$= p^*(\gamma) - \frac{1}{\gamma}(\bar{t}-t) + \frac{1}{\gamma}(\bar{t}-t)$$

$$= p^*(\gamma)$$

whereas for $t < 0$ we have $c(a^*(t)|t) = c(t|t) = 0$ and we obtain:

$$Q(t) = U(\bar{t}) - \int_t^{\bar{t}} \frac{1}{\gamma} dx + c(a^*(t)|t)$$

$$= p^*(\gamma) - \frac{1}{\gamma}(\bar{t} - t).$$

We also need to satisfy the boundary conditions, namely $U(t_*) = 0 \iff p^*(\gamma) = \frac{1}{\gamma}(\bar{t} - t_*)$. In addition, $p^*(\gamma) \leq 1$. Suppose that $\frac{1}{\gamma}(\bar{t} - \underline{t}) \leq 1$ then the constraint $p^*(\gamma) \leq 1$ does not bind and the optimal value of $t_*$ maximizes

$$\int_{t_*}^0 [p^*(\gamma) - \frac{1}{\gamma}(\bar{t} - t)] f(t) dt + \int_0^{\bar{t}} p^*(\gamma) f(t) dt = \int_{t_*}^0 [\frac{1}{\gamma}(t - t_*)] f(t) dt + \int_0^{\bar{t}} \frac{1}{\gamma}(\bar{t} - t_*) f(t) dt$$

The first-order condition is $\int_{t_*}^{\bar{t}} f(t) dt = 0$ which yields $t_* = t_0$. Together with the probability constraint with thus obtain:

$$t_* = \min \left\{ t \in [t_0, \bar{t}] : \frac{1}{\gamma}(\bar{t} - t) \leq 1 \right\}. \tag{9}$$

Therefore, whenever $(\bar{t} - t_0) \leq \gamma$ the probability constraint does not bind and $t_* = t_0$ and

$$p^*(\gamma) = \frac{1}{\gamma}(\bar{t} - t_0).$$

Else, that is when $(\bar{t} - t_0) > \gamma$, we set $p^* = 1$ and $t^*$ satisfies. Putting everything together:

$$Q^*(t) = \begin{cases} p^*(\gamma) \text{ for } t \geq 0 \\ p^*(\gamma) - \frac{1}{\gamma}(\bar{t} - t) \text{ for } t_* \leq t < 0 \\ 0 \text{ otherwise} \end{cases}$$

where $p^*(\gamma) = \min \left\{ 1, \frac{1}{\gamma}(\bar{t} - t_0) \right\}$. We have therefore established the following proposition:

**Proposition 3** *Suppose that $\frac{1}{\gamma} c(a, t) = \frac{1}{\gamma}|a - t|$. Then, the optimal mechanism is*

$$Q^*(t) = q^*(a^*(t), t) = \begin{cases} p^*(\gamma) \text{ for } a = \bar{t} \\ [p^*(\gamma) - \frac{1}{\gamma}(\bar{t} - t)]^+ \text{ for } t < 0 \end{cases} \qquad a^*(t) = \begin{cases} \bar{t} \text{ for } t \geq 0 \\ t \text{ for } t < 0 \end{cases}$$

*where $p^*(\gamma) = \min \left\{ 1, \frac{1}{\gamma}(\bar{t} - t_0) \right\}$. When $\gamma \leq \bar{t}$ the first-best is achieved, when $(\bar{t} - t_0) \geq \gamma \geq \bar{t}$ all positive types are approved with certainty while negatives are randomly approved. Finally, when $\gamma > (\bar{t} - t_0)$ all positive types are approved with $p^*(\gamma) < 1$ while are randomly approved.*

It is easy to see that the optimal mechanism can be implemented by a test that randomly assigns inputed types (here types are scores) to an approval or rejection recom-

mendation to a decision-maker. It thus does not require commitment to the decision nor type reports.

## 5.2 Quadratic cost

Suppose $\frac{1}{\gamma}c(a,t) = \frac{1}{\gamma}(a-t)^2$. The agent's payoff from truth-telling writes:

$$U(t) = \max_{t' \in T} \left( Q(t') - \int_{a \in A} \frac{1}{\gamma}(a^2 + t^2 - 2at)d\rho(a|t') \right)$$
$$= \max_{t' \in T} \left( Q(t') - \int_{a \in A} \frac{1}{\gamma}(a^2 - 2at)d\rho(a|t') \right) - \frac{1}{\gamma}t^2.$$

Choosing a report $t'$ to maximize payoff solves the following equivalent problem:

$$\tilde{U}(t) = \max_{t' \in T} \left( Q(t') - \int_{a \in A} \frac{1}{\gamma}(a^2 - 2at)d\rho(a|t') \right)$$

with modified cost function

$$\tilde{c}(a|t) \equiv \frac{1}{\gamma}(a^2 - 2at) \tag{10}$$

which is linear and thus concave in $t$.

For $a \geq t$, the derivative of $\tilde{c}$ is $-2\frac{1}{\gamma}a$ and

$$\tilde{C}(t) = 2\frac{1}{\gamma}a.$$

Then, the principal's objective becomes:

$$\int_{t_*}^{\bar{t}} Q(t)tf(t)dt = \int_{t_*}^{\bar{t}} \int_{a \in A} \left( \frac{1}{\gamma}(a^2 - 2at)t + 2\frac{1}{\gamma}a\mathbb{E}[z|z \geq t]\frac{(1 - F(t))}{f(t)} \right) d\rho(a|t)f(t)dt$$
$$= \int_{t_*}^{\bar{t}} \int_{a \in A} \left( \frac{1}{\gamma}(a^2 - 2at)tf(t) + 2\frac{1}{\gamma}a \int_t^{\bar{t}} zf(z)dz \right) d\rho(a|t)dt$$

and the corresponding relaxed problem is:

$$\max_{a \in [t,\bar{t}]} \frac{1}{\gamma}(a^2 - 2at)tf(t) + 2\frac{1}{\gamma}a \left( \int_t^{\bar{t}} zf(z)dz \right).$$

Maximizing pointwise as before we obtain the following optimal falsification target:

$$a^*(t) = t - \frac{1}{tf(t)} \left( \int_t^{\bar{t}} zf(z)dz \right).$$

We now show that whenever $F$ satisfies the monotone hazard rate property, then $a^*$ is increasing in $t$. This property will be used below to establish monotonicity of $C$ as well as to identify the type above which $Q$ reaches its maximum value.

22

**Lemma 5 (Increasing action recommendation)** *Assume F has nondecreasing hazard rate. Then, the optimal action recommendation of the relaxed problem $a^*$ is increasing in the natural score for all scores $[t_0, \bar{t}]$.*

**Verifying monotonicity** Incentive compatibility requires that $C(t)$ is increasing in $t$. Using our pointwise optimal recommendation strategy which is deterministic, $C(t)$ becomes:

$$\tilde{C}(t) = -\frac{\partial \tilde{c}(a^*(t)|t)}{\partial t} = 2\frac{1}{\gamma}a^*(t) \tag{11}$$

which is increasing in $t$ as desired under MHR because as Lemma 5 established $a^*$ is increasing. Thus the solution of the relaxed problem satisfies monotonicity.

**Identifying the growth interval** We proceed to identify the smallest type at which the approval probability reaches its highest value. Note that for $t < 0$ but very close to 0, the optimal action $a^*$ explodes to infinity so there is some $t < 0$ denoted by $t^\dagger$ such that $a^*(t^\dagger) = \bar{t}$ therefore because $a^*$ is increasing we have

$$t^\dagger = (a^*)^{-1}(\bar{t}). \tag{12}$$

By definition $t_*$ satisfies $U(t_*) = 0$. From the discussion of the case of linear costs we also know that $U(\bar{t}) = p^*(\gamma)$. Also, $U(t) = \tilde{U}(t) - \frac{1}{\gamma}t^2$ and $U(\bar{t}) + \frac{1}{\gamma}\bar{t}^2 = \tilde{U}(\bar{t})$.

Leveraging these equalities and condition 3 of Lemma 4 to express $\tilde{U}$ we obtain:

$$U(t_*) = \tilde{U}(\bar{t}) - \int_{t_*}^{\bar{t}} C(z)dz - \frac{1}{\gamma}t_*^2 = U(\bar{t}) - \int_{t_*}^{\bar{t}} C(z)dz + \frac{1}{\gamma}(\bar{t}^2 - t_*^2) = 0. \tag{13}$$

Following an analogous procedure as we did for linear costs, we can show that when $\gamma > (\bar{t} - t_0)^2$ the probability constraint does not bind and $t_* = t_0$ and using this value we can pin down $p^*(\gamma)$ because

$$U(t_0) = 0 \iff p^*(\gamma) = \int_{t_0}^{\bar{t}} 2\frac{1}{\gamma}a^*(z)dz + \frac{1}{\gamma}(\bar{t}^2 - t_0^2). \tag{14}$$

Instead, when $(\bar{t} - t_0)^2 \geq \gamma$, the probability constraint binds so $p^*(\gamma) = 1$ which together with $U(t_*) = 0$ pins down $t_*$:

$$U(t_*) = 1 - \int_{t_*}^{\bar{t}} 2\frac{1}{\gamma}a^*(z)dz + \frac{1}{\gamma}(\bar{t}^2 - t_*^2) = 0. \tag{15}$$

Note that condition 3 of Lemma 4 implies that whenever $U(t_*) = 0$ then $U(t) \geq 0$ for all $t > t_*$. Moreover by the construction of the falsification strategy for types below $t_*$ (namely no falsification and zero assignment probability) we have that the participation constraints are satisfied for all $t$. Therefore the pointwise optimal score submission is:

$$a^*(t) = \begin{cases} t \text{ for } t < t_* \\ t - \frac{1}{tf(t)} \left( \int_t^{\bar{t}} zf(z)dz \right) \text{ for } t \in [t_*, t^\dagger) \\ \bar{t} \text{ for } t^\dagger \leq t \end{cases}$$

where $t_*$ satisfies (15) or it is equal to $t_0$ and $t^\dagger$ satisfies (12).

To calculate the assignment probability we use $a^*$ and condition 3 of Lemma 4 but now scale back the cost to $c$

$$Q^*(t) = U(t) + c(a^*(t)|t)$$
$$= U(\bar{t}) + \frac{1}{\gamma}\bar{t}^2 - \int_t^{\bar{t}} C(z)dz - \frac{1}{\gamma}t^2 + \int_{a \in A} \frac{1}{\gamma}(a^2 + t^2 - 2at)d\rho^*(a|t)$$
$$= p^*(\gamma) + \frac{1}{\gamma}\bar{t}^2 - \int_{t^\dagger}^{\bar{t}} 2\frac{1}{\gamma}dz - \int_t^{t^\dagger} 2\frac{1}{\gamma}a^*(z)dz + \frac{1}{\gamma}((a^*(t))^2 - 2a^*(t)t).$$

For $\gamma > c(\bar{t}, 0)$

$$Q^*(t) = \begin{cases} 0 \text{ for } t < t_* \\ p^*(\gamma) + \frac{1}{\gamma}\bar{t}^2 - \int_{t^\dagger}^{\bar{t}} 2\frac{1}{\gamma}dz - \int_t^{t^\dagger} 2\frac{1}{\gamma}a^*(z)dz + \frac{1}{\gamma}((a^*(t))^2 - 2a^*(t)t) \text{ for } t \in [t_*, t^\dagger) \\ p^*(\gamma) \text{ for } t \in [t^\dagger, \bar{t}] \end{cases}$$

(16)

While, as anticipated earlier, whenever $\gamma \leq c(\bar{t}, 0)$ the first-best is achieved:

$$Q^*(t) = \begin{cases} 0 \text{ for } t < \bar{t} \\ 1 \text{ for } t = \bar{t}. \end{cases} \tag{17}$$

We have therefore established the following proposition:

**Proposition 4** *Suppose that $\frac{1}{\gamma}c(a, t) = \frac{1}{\gamma}(a - t)^2$ and that $F$ satisfies the monotone hazard rate property. Then the optimal mechanism is*

$$q^*(a^*(t), t) = \begin{cases} p^*(\gamma) \text{ for } t \in [t^\dagger, \bar{t}] \\ p^*(\gamma) + \frac{1}{\gamma}\bar{t}^2 - \int_{t^\dagger}^{\bar{t}} 2\frac{1}{\gamma}dz - \int_t^{t^\dagger} 2\frac{1}{\gamma}a^*(z)dz + \frac{1}{\gamma}((a^*(t))^2 - 2a^*(t)t) \text{ for } t \in [t_*, t^\dagger) \\ 0 \text{ for } t < t_* \end{cases}$$

$$a^*(t) = \begin{cases} \bar{t} \text{ for } t \in [t^\dagger, \bar{t}] \\ t - \frac{1}{tf(t)} \left( \int_t^{\bar{t}} zf(z)dz \right) \text{ for } t \in [t_*, t^\dagger) \\ t \text{ for } t < t_* \end{cases}$$

*where $t_*(v)$ satisfies (15) or it is equal to $t_0$ and $t^\dagger$ satisfies (12). When $\gamma \leq \bar{t}^2$ the first-best is achieved, when $(\bar{t} - t_0)^2 \geq \gamma \geq \bar{t}^2$ then $p^*(\gamma) = 1$ all positive types are approved with certainty while negatives are randomly approved with a probability increasing in t. Finally,*

*when $\gamma > (\bar{t} - t_0)^2$ all positive types are approved with $p^*(\gamma) = \frac{1}{\gamma}(\bar{t} - t_0)^2 < 1$ while negatives are randomly approved with a probability increasing in t.*

As in case of linear costs, it is easy to see that the optimal mechanism can be implemented by a test that randomly assigns submitted scores to an approval or rejection recommendation to a decision-maker. It thus does not require commitment nor type reports.

**Quadratic cost: uniform distribution** Suppose the $F$ is the uniform on $[-2, 1]$. Then the optimal mechanism is

$$
a^*(t) = \begin{cases} t \text{ for } t < -1 \\ \frac{3t}{2} - \frac{1}{2t} \text{ for } -1 \le t \le -\frac{1}{3} \\ 1 \text{ for } -\frac{1}{3} \le t \end{cases}
$$

and for $\gamma > 1$:

$$
Q^*(t) = \begin{cases} 0 \text{ for } t < -1 \\ p^*(\gamma) - \frac{1}{\gamma} - \frac{2}{3\gamma} + \frac{1}{\gamma}\left(\ln|-\frac{1}{3}| - \ln|a|\right) - \frac{1}{6\gamma} + \frac{6t^2}{\gamma} + \frac{1}{2\gamma t^2} - \frac{3}{\gamma} \text{ for } t \in [-1, -\frac{1}{3}) \\ p^*(\gamma) \text{ for } t \in [-\frac{1}{3}, \bar{t}) \end{cases}
$$

(18)

whereas for $\gamma \le 1$ we get the first best. The following figure depicts the optimal interim approval probability and the associated cost for two different values of $\gamma$:
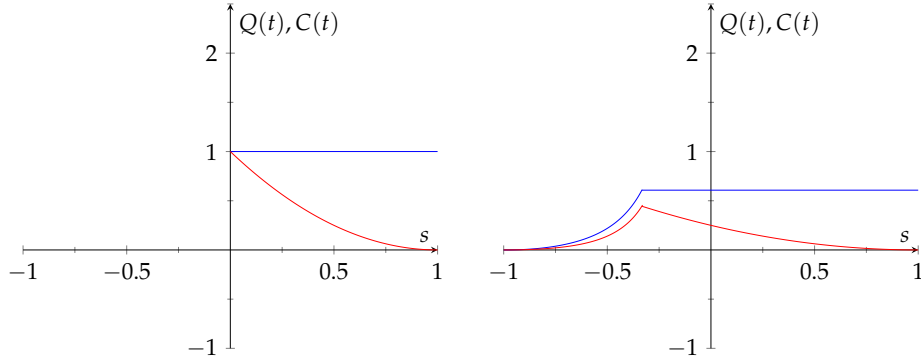


Figure 1: Left panel $\gamma = 1$;　　　Right panel $\gamma = 4$

The left panel depicts the interim approval probability for $\gamma = 1$ which is equal to the first best. The only distortion is in terms of the agent's utility loss due to falsification cost which is given by the difference $Q(t) - C(t)$. The left panel depicts the interim approval probability for $\gamma = 4$ which involves distortions because worthy types are assigned an object with probability less than 1 and unworthy types are also assigned objects. The only distortion is in terms of the agent's utility loss due to falsification cost which is given by the difference $Q(t) - C(t)$.

# A  Proof of Lemma 2

Consider an IC DRM mechanism $\nu$. Let $\tilde{q}$ satisfy

$$\sum_{q \in D} \nu(q, a \mid t) q(x|a) = \tilde{q}(x|a) \sum_{q \in D} \nu(q, a \mid t). \tag{19}$$

Note that $\tilde{q}$ is a valid score-based decision rule, that is, $\tilde{q}$ maps $A$ to $\Delta(X)$. First, $0 \leq \tilde{q}(x|s) \leq 1$ because

$$\tilde{q}(x|a) = \frac{\sum_{q \in D} \nu(q, a \mid t) q(x|a)}{\sum_{q \in D} \nu(q, a \mid t)}$$

and $q(x|a) \in [0, 1]$. Moreover,

$$\sum_{x \in X} \tilde{q}(x|a) = \frac{\sum_{x \in X} \sum_{q \in D} \nu(q, a \mid t) q(x|a)}{\sum_{q \in D} \nu(q, a \mid t)} = \frac{\sum_{q \in D} \nu(q, a \mid t) \underbrace{\sum_{x \in X} q(x|a)}_{=1}}{\sum_{q \in D} \nu(q, a \mid t)} = 1.$$

Define a new mechanism $\mu = \alpha \circ \rho$ where

$$\alpha(t') = \delta_{\tilde{q}} \text{ and } \rho(a \mid t') \equiv \sum_{q \in D} \nu(q, a \mid t) \quad \forall t' \in T. \tag{20}$$

Note that $\rho$ is the marginal of the mechanism $\nu$ over $D$. Observe that the agent's payoff is the same under $\nu$ and $\mu$ because, conditional on each possible report $t' \in T$, evidence request $a \in A$ and for each $x \in X$ we have:

$$\sum_{q \in D} \nu(q, a \mid t) q(x|a) [v(x, t) - c(a, t)]$$
$$= \tilde{q}(x|a) \sum_{q \in D} \nu(q, a \mid t) [v(x, t) - c(a, t)]$$
$$= \tilde{q}(x|a) \rho(a \mid t') [v(x, t) - c(a, t)]$$
$$= \mu(\tilde{q}, a \mid t') [v(x, t) - c(a, t)]$$

where the first quality uses (19) and (20) and the second and third equalities use (20) and the definition of $\mu$. Summing up over all $x \in X$ and $a \in A$ we obtain:

$$U(t, t', \nu) = \sum_{x \in X} \sum_{a \in A} \sum_{q \in D} \nu(q, a \mid t) q(x|a) [v(x, t) - c(a, t)]$$
$$= \sum_{x \in X} \sum_{a \in A} \mu(\tilde{q}, a \mid t') [v(x, t) - c(a, t)]$$
$$= U(t, t', \mu)$$

where $U(t, t', \cdot)$ is defined in (1).

# B Proof of Proposition 2

Fix an incentive compatible DRM that has a deterministic score recommendation rule. Consider two types $t$ and $t'$ with $\rho(\cdot|t) = \rho(\cdot|t') = \delta_a$. Suppose that $q_\mu(x|a,t) \neq q_\mu(x|a,t')$ for some $x$. Truth-telling implies:

$$\sum_x q_\mu(x|a,t)\beta(a,t)v(x) \geq \sum_x q_\mu(x|a,t')\beta(a,t)v(x) \iff \sum_x q_\mu(x|a,t)v(x) \geq \sum_x q_\mu(x|a,t')v(x)$$

where the equivalence follows because $\beta(a,t) > 0$. Analogously we obtain:

$$\sum_x q_\mu(x|a,t')\beta(a,t')v(x) \geq \sum_x q_\mu(x|a,t)\beta(a,t')v(x) \iff \sum_x q_\mu(x|a,t')v(x) \geq \sum_x q_\mu(x|a,t)v(x)$$

Combining results to

$$\sum_x q_\mu(x|a,t')v(x) = \sum_x q_\mu(x|a,t)v(x) \iff \sum_x [q_\mu(x|a,t') - q_\mu(x|a,t)]v(x) = 0$$

The last inequality implies that the vectors $q_\mu(x|a,t) - q_\mu(x|a,t')$ and $v(x)$ are orthogonal. We also know $\sum_x [q_\mu(x|a,t') - q_\mu(x|a,t)]\mathbf{1} = \mathbf{0}$. Letting $\lambda(x) = q_\mu(x|a,t) - q_\mu(x|a,t')$, we summarize the above two observations as follows:

$$\sum_x \lambda(x)\mathbf{1} = 0 \qquad \sum_x \lambda(x)v(x) = 0. \tag{21}$$

Now consider the principal and suppose

$$\sum_x q_\mu(x|a,t)[w(a,t)r(x,a) + y(a,t)v(x) + \ell(a,t)] \geq \sum_x q_\mu(x|a,t')[w(a,t)r(x,a) + y(a,t)v(x) + \ell(a,t)] \iff$$

$$\sum_x q_\mu(x|a,t)[w(a,t)r(x,a) + y(a,t)v(x)] \geq \sum_x q_\mu(x|a,t')[w(a,t)r(x,a) + y(a,t)v(x)] \iff$$

$$\sum_x [q_\mu(x|a,t) - q_\mu(x|a,t')][w(a,t)r(x,a) + y(a,t)v(x)] \geq 0$$

$$\sum_x [q_\mu(x|a,t) - q_\mu(x|a,t')]\left[r(x,a) + \frac{y(a,t)}{w(a,t)}v(x)\right] \geq 0$$

$$\sum_x [q_\mu(x|a,t) - q_\mu(x|a,t')][r(x,a) + z(a,t)v(x)] \geq 0$$

where the first implication uses the first equality in (21), the second is a simple rewriting, the third uses the fact that $w(a,t) > 0$ and in the fourth we let $z(a,t) \equiv \frac{y(a,t)}{w(a,t)}$. The above final inequality writes as

$$\sum_x [\lambda(x)r(x,a) + z(a,t)\lambda(x)v(x)] \geq 0 \iff \sum_x \lambda(x)r(x,a) \geq 0$$

where the equivalence uses (21). The final inequality implies that the principal ranks the lotteries $q_\mu(\cdot|a,t)$ and $q_\mu(\cdot|a,t')$ in the same way no matter the agent's type realization. And because agent types are indifferent we can offer the principal's preferred lottery that only depends on $a$ and we call it $q(\cdot|a)$.

## C  Proof of Lemma 3

Consider a type $t$ and all scores on the support of $t$'s falsification strategy. Denote them $A(t)$. Suppose, for simplicity, $A(t)$ is finite. The idea extends to arbitrary supports. Rank its elements from smallest to largest and label so that

$$a_1 < a_2 < \cdots < a_n$$

relabel the corresponding assignment probabilities so that

$$q_i = q(a_i, t)$$

and falsification requests so that

$$\rho_i = \rho(a_i, t)$$

Trivially, this relabelling by construction ensures the payoff to $t$ and to mimicking $t'$ is the same. To see this note:

$$Q(t) = \sum_{i=1}^n \rho_i q_i = \sum_{i=1}^n \rho(a_i, t)a(a_i, t) = \sum_{a \in A(t)} \rho(a, t)a(s, t) \text{ and}$$

$$C(t) = \sum_{i=1}^n \rho_i c(a_i, t) = \sum_{i=1}^n \rho(a_i, t)c(a_i, t) = \sum_{a \in A(t)} \rho(a, t)c(a, t) \text{ and}$$

$$C(t \mid t') = \sum_{i=1}^n \rho_i c(a_i, t') = \sum_{i=1}^n \rho(a_i, t)c(a_i, t') = \sum_{a \in A(t)} \rho(a, t)c(a, t').$$

If $q_1 \geq 1$, then because $q_1 \leq 1$ this is only possible when $q_i = 1 \forall i \in \{1, \ldots, n\}$ hence the approval probability is increasing in $s$ and there is nothing to prove. If the approval probability is not increasing, then

$$q_1 < 1. \tag{22}$$

We construct an increasing payoff-equivalent approval probability as follows:

$$\tilde{q}_i = c(a_i, t) \forall i < n \text{ and } \tilde{q}_n = \frac{\rho_n q_n + \sum_{i \neq n} q_i \rho_i - \sum_{i \neq n} c(a_i, t)_i \rho_i}{\rho_n} \tag{23}$$

$$\rho_n \tilde{q}_n + \sum_{i \neq n} c(a_i, t)_i \rho_i = \rho_n q_n + \sum_{i \neq n} q_i \rho_i \geq \sum_{i=1}^n \rho(a_i, t)c(a_i, t) = \sum_{a \in A(t)} \rho(a, t)c(a, t) \tag{24}$$

where the last inequality follows from obedience that requires

$$q_i - c(a_i, t) \geq 0 \forall i.$$

Then, for (24) to hold, it must be the case that $\tilde{q}_n \geq c(a_n, t)$ which ensures obedience and that $\tilde{q}$ is increasing given that $c(a, t)$ is increasing in $s$ which implies that

$$\tilde{q}_n \geq c(a_n, t) \geq c(s_{n-1}, t) = \tilde{q}_{n-1} \geq \ldots c(s_1, t) = \tilde{q}_1.$$

The problem is that it is possible that $\tilde{q}_n > 1$. In that case we modify the construction as follows: Set $\hat{a}_n = 1$ and assigning the difference to lower scores:

$$d_n \equiv \frac{\rho_n q_n + \sum_{i \neq n} q_i \rho_i - \sum_{i \neq n} c(a_i, t)_i \rho_i}{\rho_n} - 1.$$

We use this equivalent expression:

$$\rho_n d_n = \rho_n q_n + \sum_{i \neq n} q_i \rho_i - \sum_{i \neq n} c(a_i, t)_i \rho_i - \rho_n$$

to increase $\tilde{q}_{n-1}$ to $\hat{a}_{n-1} = \tilde{q}_{n-1} + d_{n-1}$ where $d_{n-1}$ satisfies

$$d_{n-1} \rho_{n-1} = d_n \rho_n.$$

If $\hat{a}_{n-1} \leq 1$. We are done. Otherwise, continue in this way. At some point we will stop because (22) implies that we cannot have all 1's. The resulting assignment is increasing, satisfies OB because all assignment probabilities assigned to a score are by construction higher than the cost type $t$ incurs to generate that score. Also,

$$\rho_n + d_{n-1} \rho_{n-1} + \sum_{i \neq n} \rho_i c(s_t | t)$$
$$= \rho_n + d_n \rho_n + \sum_{i \neq n} \rho_i c(s_t | t)$$
$$= \rho_n + \rho_n q_n + \sum_{i \neq n} q_i \rho_i - \sum_{i \neq n} c(a_i, t)_i \rho_i - \rho_n + \sum_{i \neq n} \rho_i c(s_i, t)$$
$$= \rho_n q_n + \sum_{i \neq n} q_i \rho_i - \sum_{i \neq n} c(a_i, t) \rho_i + \sum_{i \neq n} \rho_i c(s_i, t)$$
$$= Q(t).$$

Hence the modification results to the same expected approval probability and the same falsification costs for all types and it is payoff equivalent for the agent and the designer.

# D   Proof of Lemma 5

Differentiating $a^*$ results to:

$$\frac{\partial a^*(t)}{\partial t} = 2 + \frac{1}{t^2 f(t)}\left(\int_t^{\bar{t}} z f(z)dz\right) + \frac{f'(t)}{t(f(t))^2}\left(\int_t^{\bar{t}} z f(z)dz\right)$$

$$\geq 2 + \frac{1}{t^2 f(t)}\left(\int_t^{\bar{t}} z f(z)dz\right) + \frac{f'(t)}{t(f(t))^2}\left(\int_t^{\bar{t}} t f(z)dz\right)$$

$$= 2 + \frac{1}{t^2 f(t)}\left(\int_t^{\bar{t}} z f(z)dz\right) + f'(t)\frac{1 - F(t)}{(f(t))^2}$$

where the inequality follows because $t \leq z$. Now for $t \in [t_0, \bar{t}]$, $\mathbb{E}[z|z \geq t] \geq 0$, which implies that only the last term, namely $f'(t)\frac{1-F(t)}{(f(t))^2}$ could be negative.

Recall that the derivative of the usual virtual valuation, namely $J(t) = t - \frac{1-F(t)}{f(t)}$ is:

$$\frac{\partial J(t)}{\partial t} = 2 + f'(t)\frac{1 - F(t)}{(f(t))^2}.$$

A sufficient condition for $J$ to be increasing is that the distribution has nondecreasing hazard rate. More generally, whenever the usual virtual valuation is increasing, the optimal recommended action is increasing. This is because the derivative is $\frac{\partial a^*(t)}{\partial t} = \frac{\partial J(t)}{\partial t} + \frac{1}{t^2 f(t)}\left(\int_t^{\bar{t}} z f(z)dz\right)$ and the term we are adding is positive, hence increasing $J$ (which is ensured by MHR) suffices for $a^*$ to be increasing.

# References

AKBARPOUR, M., P. DWORCZAK, AND F. YANG (2023): "Comparison of Screening Devices," *Available at SSRN 4456198*.

BALL, I. (2022): "Scoring Strategic Agents," *Working Paper*.

BEN-PORATH, E., E. DEKEL, AND B. L. LIPMAN (2019): "Mechanisms with evidence: Commitment and robustness," *Econometrica*, 87, 529–566.

BEN-PORATH, E. AND B. L. LIPMAN (2012): "Implementation with partial provability," *Journal of Economic Theory*, 147, 1689–1724.

BJERRE-NIELSEN, A., L. S. CHRISTENSEN, M. H. GANDIL, AND H. H. SIEVERTSEN (2023): "Playing the system: address manipulation and access to schools," .

BOLTON, L. (2018): "Manipulation of the Waitlist Priority of the Organ Allocation System through the Escalation of Medical Therapies," *OPTN/UNOS EthicsCommittee*.

BRAVERMAN, M. AND S. GARG (2020): "The role of randomness and noise in strategic classification," *arXiv preprint arXiv:2005.08377*.

BULL, J. (2008a): "Costly evidence production and the limits of verifiability," *The BE Journal of Theoretical Economics*, 8.

——— (2008b): "Mechanism design with moderate evidence cost," *The BE Journal of Theoretical Economics*, 8.

BULL, J. AND J. WATSON (2007): "Hard evidence and mechanism design," *Games and Economic Behavior*, 58, 75–93.

CHAKRAVARTY, S. AND T. R. KAPLAN (2013): "Optimal allocation without transfer payments," *Games and Economic Behavior*, 77, 1–20.

CONDORELLI, D. (2012): "What money can't buy: Efficient mechanism design with costly signals." *Games Econ. Behav.*, 75, 613–624.

CROCKER, K. AND J. MORGAN (1998): "Is Honesty the Best Policy? Curtailing Insurance Fraud through Optimal Incentive Contracts," *Journal of Political Economy*, 106, 355–375.

DENECKERE, R. AND S. SEVERINOV (2008): "Mechanism Design with Partial State Verifiability," *Games and Economic Behavior*, 64, 487–513, [327].

———— (2022): "Signalling, screening and costly misrepresentation," *Canadian Journal of Economics/Revue canadienne d'économique*, 55, 1334–1370.

DWORCZAK, P. (2022): "Equity-efficiency trade-off in quasi-linear environments," Working paper.

FORGES, F. AND F. KOESSLER (2005): "Communication equilibria with partially verifiable types," *Journal of Mathematical Economics*, 41, 793–811.

FRANKEL, A. AND N. KARTIK (2019): "Muddled information," *Journal of Political Economy*, 127, 1739–1776.

———— (2021): "Improving information from manipulable data," *Journal of the European Economic Association*, jvab017.

GLAZER, J. AND A. RUBINSTEIN (2004): "On Optimal Rules of Persuasion," *Econometrica*, 72, 1715–1736.

———— (2006): "A study in the pragmatics of persuasion: a game theoretical approach," *Theoretical Economics*, 1, 395–410.

GREEN, J. R. AND J.-J. LAFFONT (1986): "Partially verifiable information and mechanism design," *The Review of Economic Studies*, 53, 447–456.

HARDT, M., N. MEGIDDO, C. PAPADIMITRIOU, AND M. WOOTTERS (2016): "Strategic classification," in *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

HART, S., I. KREMER, AND M. PERRY (2017): "Evidence games: Truth and commitment," *American Economic Review*, 107, 690–713.

HARTLINE, J. D. AND T. ROUGHGARDEN (2008): "Optimal mechanism design and money burning," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 75–84.

HU, L., N. IMMORLICA, AND J. W. VAUGHAN (2019): "The disparate effects of strategic manipulation," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

KARTIK, N. AND O. TERCIEUX (2012): "Implementation with evidence," *Theoretical Economics*, 7, 323–355.

KEPHART, A. AND V. CONITZER (2016): "The revelation principle for mechanism design with reporting costs," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, 85–102.

LACKER, J. M. AND J. A. WEINBERG (1989): "Optimal Contracts with Costly State Falsification," *Journal of Political Economy*, 97, 1345–1363.

LI, Y. AND X. QIU (2023): "Screening Signal-Manipulating Agents via Contests," .

MAGGI, G. AND A. RODRIGUEZ-CLARE (1995): "Costly Distortion of Information in Agency Problems," *RAND Journal of Economics*, 26, 675–689.

MASKIN, E. (1999): "Nash equilibrium and welfare optimality," *The Review of Economic Studies*, 66, 23–38.

MCMICHAEL, B. J. (2022): "Stealing Organs?" *Indiana Law Journal*, 97, 135.

MUSSA, M. AND S. ROSEN (1978): "Monopoly and product quality," *Journal of Economic Theory*, 18, 301–317.

MYERSON, R. B. (1982): "Optimal coordination mechanisms in generalized principal-agent problems," *Journal of Mathematical Economics*, 10, 67–81.

PATHAK, P. A. AND T. SÖNMEZ (2008): "Leveling the playing field: Sincere and sophisticated players in the Boston mechanism," *American Economic Review*, 98, 1636–1652.

PEREZ-RICHET, E. AND V. SKRETA (2022): "Test design under falsification," *Econometrica*, 90, 1109–1142.

——— (2023): "Fraud-proof non-market allocation mechanisms," .

SCHWEIGHOFER-KODRITSCH, S. AND R. STRAUSZ (2022): "Principled Mechanism Design with Evidence," .

SEVERINOV, S. AND T. Y.-C. TAM (2019): "Screening Under Fixed Cost of Misrepresentation," .

SHER, I. (2011): "Credibility and determinism in a game of persuasion," *Games and Economic Behavior*, 71, 409–419.

SPENCE, M. (1978): "Job market signaling," in *Uncertainty in economics*, Elsevier, 281–306.

TAN, T. Y. (2023): "Price Discrimination with Manipulable Observables," *Available at SSRN 4480623*.

TANG, Z., J. ZHANG, AND K. ZHANG (2023): "What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective," *ACM Computing Surveys*, 55, 1–37, online publication date: 31-Dec-2024.