The Optimal Design of Public Recognition Schemes*

Martin Vaeth†

Preliminary – comments very welcome

Latest version here

November 14, 2024

Abstract

This paper studies the optimal design of public recognition schemes to incentivize agents who care about their social image – the public's belief about their private type, such as ability or prosociality. We allow public recognition schemes to take the form of any signal structure, employing an information design approach. When agents are risk neutral over image, we show optimality of monotone partitional recognition schemes. If agents are risk averse over image, it may be optimal to maintain full privacy and not screen the agent. Our paper highlights a difference between public recognition schemes and monetary incentive schemes, where screening remains optimal even under risk aversion.

*I would like to thank Fedor Sandomirskiy for guidance and, for helpful discussions and feedback, Roland Bénabou, Laura Doval, Andreas Kleiner, Alessandro Lizzeri, Xiaosheng Mu, Pietro Ortoleva, Wolfgang Pesendorfer, and Ali Shourideh. Financial support was provided by The William S. Dietrich II Economic Theory Center. All errors are my own.

†Princeton University. mvaeth@princeton.edu

1 Introduction

Economics typically focuses on *monetary* incentives. However, a growing body of evidence shows that *social image* is a powerful motivator beyond material rewards (Bursztyn and Jensen, 2017). Because individuals care about how they are perceived by others, public recognition schemes – which signal agents' characteristics through conveying information about their behavior – can be used to incentivize desired actions. For instance, organizations implement "employee of the month" programs that signal high performance and incentivize effort. Schools employ grading systems that signal ability to the job market, motivating human capital accumulation. Charities publicly acknowledge donors, signaling prosociality and incentivizing donations. This paper studies how public recognition schemes can be optimally designed to leverage these image concerns and promote desired behaviors.

While the optimal design of public recognition schemes has been studied across various contexts,[1] the existing literature leaves two fundamental questions open.

First, what type of signal structures emerge as optimal when the principal is not constrained to specific forms? Previous work assumes particular forms of signal structures. These constraints may overlook other potentially optimal designs, so understanding the unconstrained optimal signal structures is crucial for both theory and practice.

Second, how do agents' preferences over their image affect the optimal design of recognition schemes? Specifically, prior studies assume that agents are risk neutral with respect to their image – they care linearly about the public's expectation about their type. However, if agents are risk averse regarding their image, then revealing information imposes a social welfare cost. Therefore, it is important to understand optimal public recognition schemes when agents exhibit risk aversion over their image.

In this paper, we make progress on these questions by leveraging majorization and information design tools. This approach allows us to study the optimal design problem without imposing any restrictions on the types of information structures available to the principal. By doing so, we provide a more general and robust framework that can accommodate a variety of settings and agent preferences.

To address these questions, we consider the following model. An agent has a private type θ – such as ability or degree of prosociality – drawn from a known distribution over the interval $\Theta = [\underline{\theta}, \overline{\theta}]$. Upon observing her realized type, the agent chooses an action $a \in \mathbb{R}_+$, incurring a cost that increases with higher actions and decreases with her type. The public, however, only receives a signal that is correlated with the agent's action. The agent values the public's belief about her type – her *image*. The principal designs the signal structure, referred to as a public recognition scheme,

[1] Previous literature has studied the optimal design of: privacy under prosocial signaling (Bénabou and Tirole, 2006; Bénabou and Tirole, 2011; Ali and Bénabou, 2020), performance evaluation under worker career concerns (Holmström, 1999; Dewatripont et al., 1999), quality certification under endogenous quality provision (Albano and Lizzeri, 2001; Zapechelnyuk, 2020), grading schemes under job-market signaling (Dubey and Geanakoplos, 2014; Zubrickas, 2015; Onuchic and Ray, 2023), monopolistic pricing under conspicuous consumption (Rayo, 2013), and taxation under welfare stigma (Blumkin et al., 2015).

with the objective of incentivizing the agent to take higher actions. We investigate the optimal signal design when the signal is the sole instrument available to the principal for influencing the agent's behavior.

Our first result concerns the case when the agent is risk neutral with regard to image and cares about the public's expectation about her type. The resulting design problem is tractable despite its high dimensionality, and a simple class of signal structures – *monotone partitional signals* – emerges as optimal. Monotone partitional signals resemble grading or rating schemes and are often used in practice.

To understand why these signals emerge as optimal, we need to understand the set of feasible image payoffs within our model. The set of feasible image payoffs was characterized by Saeedi and Shourideh (2024) in the context of rating design under moral hazard. By employing the majorization tools from Kleiner et al. (2021), we provide a short alternative proof of this characterization, demonstrating that any second-order expectation can be generated by publicly randomizing over monotone partitional signals. Leveraging this connection to majorization, we generalize previous results from the literature on public recognition schemes. Specifically, we show that in quasi-linear settings, a monotone partitional signal structure is optimal – that is, the principal does not need to randomize – and we analyze when the optimal signal is fully revealing or involves pooling over certain intervals of types. Finally, our proof extends to higher-order expectations, showing they have the same characterization as second-order expectations.

In the second part of our paper, we relax the risk neutrality assumption that is ubiquitous in previous work. Prior studies implicitly assume that agents are risk-neutral with respect to their image by assuming they care only about the public's expected belief about their type.

While risk neutrality is a natural starting point, there are two reasons to consider risk aversion over image as important. First, agents may have an intrinsic valuation of their image due to psychological factors, leading to risk-averse behavior. Butera et al. (2022) provide evidence for risk aversion over prosocial image, which supports this interpretation. Second, risk aversion over image can be induced from risk aversion over other goods when image has instrumental value. For example, if an agent values the public's belief about her ability because her attainable wage in the labor market depends on her expected ability, then risk aversion over income translates into risk aversion over image.

Our second main result is that under risk aversion over image, full privacy may become optimal, leading the principal to forego the possibility of screening the agent. We illustrate this result in a two-type model and contrast it with monetary incentive schemes, where the principal always employs some degree of screening to provide incentives.

The intuition behind this result is as follows. First of all, risk aversion over image makes revealing information socially costly due to Jensen's inequality: disclosing information creates a mean-preserving spread in the distribution of beliefs, lowering aggregate welfare. The principal internalizes this cost either because they directly care about the agent's well-being or because of the agent's participation constraint. Consequently, risk aversion pushes towards greater privacy.

While this force resembles the familiar efficiency-equity trade-off in the optimal taxation literature following Mirrlees (1971), we demonstrate a qualitative difference between image and standard goods like money. In traditional screening problems involving monetary transfers, there is a trade-off between providing incentives (through unequal allocations) and mitigating welfare losses from risk aversion (favoring more equal allocations). Because, for small incentives, the loss from risk aversion is of second order, some incentive provision is always optimal; complete equality (or "communism") is suboptimal.

However, when it comes to image, the push towards equality can be so strong that full privacy – the only public recognition scheme that equalizes image across agents – becomes optimal. We show that the aggregate welfare loss and the gain from providing incentives are of the same order because to induce small incentives, the principal has to resort to random signal structures, and consequently random allocations of image. Consequently, it may be optimal not to reveal information at all because any level of information would impose a greater welfare loss than benefit through providing incentives.

This result has implications for the economics of privacy, reviewed in Acquisti et al. (2016). While some authors have argued for publicity as a means to incentivize socially desirable behavior and provide economically useful information (Posner, 1978), a lack of privacy can lead to data misuse by firms or governments (for the latter, see Tirole, 2021), wasteful signaling (Daughety and Reinganum, 2010), and suppression of learning about societal preferences (Ali and Bénabou, 2020). As pointed out by Butera et al. (2022), there is also a more immediate welfare cost of publicity stemming from risk aversion over social image. We show that this force is so strong that complete privacy may be optimal despite publicity incentivizing desirable behavior.

This paper is structured as follows. Section 2 introduces the model. Section 3 studies the model when agents are risk neutral over image. Section 4 studies the model when agents are risk averse over image.

2 Baseline Model

Let θ be an agent's private type drawn from $\Theta = [\underline{\theta}, \overline{\theta}]$ according to some CDF $F : \Theta \to \mathbb{R}$ with continuous and strictly positive derivative f. By θ we will denote a generic element of Θ and we use bold symbols, such as $\boldsymbol{\theta}$, for random variables. The agent incurs linear cost a/θ for action $a \in \mathbb{R}_{\geq 0}$, say effort; that is, higher types have smaller costs. The principal wants to maximize the aggregate action $\int a(\theta) dF(\theta)$. Without loss, we restrict attention to direct mechanisms.[2] The principal recommends an action schedule $a(\theta)$. To implement this action schedule, the principal does not use monetary transfers but instead makes use of image motives by committing to revealing a signal about the reported type.

As mentioned in the introduction, the model has a number of different interpretations, such

[2]One can show that the usual taxation principle holds: we can let the agent choose the action and the principal chooses a signal structure about the chosen action. The principal can pool unchosen actions with the outside option, which gives the worst reputation, see Rayo (2013).

as a firm designing a performance evaluation scheme to incentive effort when workers have career concerns, a school designing a grading schemes to incentive human capital accumulation under job-market signaling, or, more abstractly, the optimal degree of privacy about agent's public good contributions when they engage in prosocial signaling.

Signal Structures and Image Payoff A version of this model was considered among others by Rayo (2013), who assumed that the principal is restricted to partitional signal structures. We do not impose any assumptions on the nature of signals and allow for arbitrary signal structures. A *signal structure* (S, σ) consists of the set of signals S, and a map $\sigma\colon \Theta \to \Delta(S)$ assigning a distribution of signals to each realized state.[3] Equivalently, a signal structure can be modeled as a Bayes-consistent distribution $\tau \in \Delta(\Delta(\Theta))$ over posteriors $\pi \in \Delta(\Theta)$ (Kamenica and Gentzkow, 2011). The distribution τ is Bayes consistent if $\mathbb{E}_\tau[\pi] = \mu$, where μ is the measure $\mu \in \Delta(\Theta)$ associated with CDF F. In general, agents can value the belief π held by the public as $v(\pi)$, $v\colon \Delta(\Theta) \to \mathbb{R}$. Because the agent knows their type θ, they need to condition on $\boldsymbol{\theta} = \theta$ to form their expected image payoff. We say the payoff $V : \Theta \to \mathbb{R}$ is *generated* by a signal structure $\tau \in \Delta(\Delta(\Theta))$ if $V(\theta)$ corresponds to the expected value $v(\pi)$ of the posterior π, conditioning on the true state being θ, that is,

$$V(\theta) = \mathbb{E}_\tau[v(\pi)|\boldsymbol{\theta} = \theta].$$

We say $V : \Theta \to \mathbb{R}$ is *feasible* if it is generated by some signal structure τ.

Agent's Utility The utility $\tilde{U}(\theta, \theta')$ of agent θ from reporting θ' is

$$\tilde{U}(\theta, \theta') = V(\theta') - \frac{1}{\theta} a(\theta').$$

Equivalently, we can define the agent θ as choosing θ' to maximize utility $U(\theta, \theta') = \theta\tilde{U}(\theta, \theta')$, that is,

$$U(\theta, \theta') = \theta V(\theta') - a(\theta').$$

Instead of letting the principal choose a signal structure, we write their problem as choosing the second-order expectation $V(\theta)$ subject to feasibility, that is, $V(\theta)$ can be generated by some signal structure.

Principal's Problem The principal maximizes the expected action subject to incentive compatibility (IC), voluntary participation or individual rationality (IR), and feasibility (F) of the second-order expectation $V(\theta)$. To formulate the participation constraint, we assume that upon not participating, the agent obtains the minimal second-order expectation $\underline{\theta}$, which is the most

[3] To ensure that conditional expectations are well-defined, we require that the set of signals S is a Polish space endowed with its Borel sigma-algebra and $\sigma\colon \Theta \to \Delta(S)$ is such that for all measurable $A \subseteq S$, $\sigma(A, \theta)$ is measurable in θ.

favorable to the principal. Later, we will show that this assumption is without loss.

$$\max_{a,V:\Theta\to\mathbb{R}} \int_{\underline{\theta}}^{\overline{\theta}} a(\theta) dF(\theta) \tag{P}$$

s.t.

$$\theta V(\theta) - a(\theta) \geq \theta V(\theta') - a(\theta') \qquad \forall \theta, \theta' \tag{IC}$$

$$\theta V(\theta) - a(\theta) \geq \theta \cdot \underline{\theta} \qquad \forall \theta \tag{IR}$$

$$V \text{ is feasible.} \tag{F}$$

The principal's problem (P) looks similar to a standard screening problem where the principal allocates $V(\theta)$, usually money or the allocation probability of a good, to implement action schedule $a(\theta)$. However, the screening problem has two peculiarities. First, is has a type-dependent participation constraint (IR). Second, it has a non-standard feasibility constraint (F). We first deal with (IR) and simplify the principal's problem, before we turn to the feasibility constraint in section 3.

Rewriting the Principal's Problem Suppose (IR) holds for $\underline{\theta}$. By (IC) and the envelope theorem, the derivative of the left-hand side of (IR) in θ is $V(\theta)$. The derivative of the right-hand side in θ is $\underline{\theta}$. Any feasible $V(\theta)$ satisfies $V(\theta) \in [\underline{\theta}, \overline{\theta}]$, thus if (IR) holds for $\underline{\theta}$, it holds for all θ. Thus, we can rewrite (IR) as

$$\underline{\theta} \cdot \underline{\theta} \qquad \forall \theta \tag{IR'}$$

By supermodularity of the agent's utility in θ and $V(\theta)$, (IC) implies that $V(\theta)$ is *monotone*, that is $\theta' > \theta$ implies $V(\theta') \geq V(\theta)$. Following standard arguments we can rewrite the principal's problem as a linear maximization problem P' in $V(\theta)$ subject to $V(\theta)$ being monotone and feasible.

$$\max_{V:\Theta\to\mathbb{R}} \int V(\theta) \left(\theta - \frac{1 - F(\theta)}{f(\theta)} \right) dF(\theta) \tag{P'}$$

s.t.

$$V(\theta) \leq V(\theta') \qquad \forall \theta < \theta' \tag{MON}$$

$$V \text{ is feasible.} \tag{F}$$

The principal's problem reduces to a linear optimization problem over feasible, monotone V. The crucial question is what monotone $V(\theta)$ are feasible.

3 Risk Neutrality over Image

In this section, we consider the special case where the value of a posterior equals the expected type,

$$v(\pi) = \mathbb{E}_\pi[\boldsymbol{\theta}],$$

as assumed by virtually all of the previous literature. In this case, $V(\theta)$ equals the second-order expectation $P(\theta)$, defined as follows. A signal structure (S, σ) defines a joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{s}$: for each realized state $\boldsymbol{\theta} = \theta$, a random signal $\mathbf{s}$ is drawn from S according to the distribution $\sigma(\theta)$. We define the *second-order expectation* generated by signal structure (S, σ) as

$$P(\theta) := \mathbb{E}[\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{s}]|\boldsymbol{\theta} = \theta].$$

We say $P(\theta)\colon \Theta \to \mathbb{R}$ is *feasible* if it is generated by some signal structure (S, σ). For clarity of exposition, we use $P(\theta)$ instead of $V(\theta)$ in this section.

Saeedi and Shourideh (2022) study second-order expectations in the context of moral hazard and characterize the feasible, monotone $P(\theta)$. To state their result, we introduce the following notation.

For a subset $\Omega \subset \mathbb{R}$, let $L^1(\Omega)$ denote the real-valued, integrable, and right-continuous functions defined on Ω. For monotone functions $P \in L^1(\Theta)$, we say that P is *majorized by the identity function with respect to* F, denoted by $P \prec_F \text{id}$, if

$$\begin{aligned} \int_{\hat{\theta}}^{\overline{\theta}} P(\theta) dF(\theta) &\leq \int_{\hat{\theta}}^{\overline{\theta}} \theta dF(\theta) \quad \forall \hat{\theta} \in (\underline{\theta}, \overline{\theta}], \\ \int_{\underline{\theta}}^{\overline{\theta}} P(\theta) dF(\theta) &= \int_{\underline{\theta}}^{\overline{\theta}} \theta dF(\theta). \end{aligned} \tag{1}$$

Note that the usual majorization constraint is written with integrals with respect to the Lebesgue measure, while the integrals in this generalized majorization relation are with respect to the measure with CDF F. Under a continuous CDF F, one can rewrite (MAJ) as a standard majorization constraint by going into quantile space through a change of variables. This generalized majorization constraint is appropriate for the following result due to Saeedi and Shourideh (2022) because it also holds for discrete distributions (which we do not show in the present version of this draft).

Theorem 1 (Saeedi and Shourideh, 2022)**.** *Monotone $P \in L^1(\Theta)$ are feasible if and only if $P \prec_F \text{id}$, that is, P is majorized by the identity function with respect to F.*

Saeedi and Shourideh (2022) prove their results using an induction argument for finite distributions of θ and an approximation argument for continuous distributions. Here, we present a short proof of their result based on extreme point and majorization techniques, building on a result from Kleiner et al. (2021). In addition to showing a connection between this fundamental result and results in majorization, our proof delivers additional insights relevant to our screening problem, see the following subsection 3.1.

Before we show our proof, we state two preliminary result about majorization and define monotone partitional signal structures. The first statement is a known result and will help us show that any feasible monotone $P(\theta)$ satisfies $P \prec_F \text{id}$. For completeness, we include a proof in the Appendix.

Lemma 1. *For monotone $P \in L^1(\Theta)$, $P \prec_F \mathrm{id}$ if and only if the distribution of $P(\boldsymbol{\theta})$ is a mean-preserving contraction of the distribution of $\boldsymbol{\theta}$.*

For the proof of Lemma 1, note that mean-preserving contraction has a simple characterization in terms of integral inequalities similar to (1) (Hardy et al., 1934), although these integrals are in terms of the CDFs of the two random variables. However, by going into quantile space and using the monotonicity of $P(\theta)$, one can translate (1) into the integral characterization of mean-preserving contraction.

Lemma 2. *Let $g \in L^1(\Theta)$ be monotone. Then, P is an extreme point of*

$$\mathrm{MPS}_F(g) := \{P \in L^1(\Theta) \mid P \text{ is monotone and } P \prec_F g\}$$

if and only if there exists a collection of disjoint intervals $[\underline{\theta}_i, \overline{\theta}_i)$ indexed by $i \in I$ such that for almost every $\theta \in [\underline{\theta}, \overline{\theta}]$,

$$P(\theta) = \begin{cases} g(\theta) & \text{if } \theta \notin \bigcup_{i\in I}[\underline{\theta}_i, \overline{\theta}_i), \\ \frac{\int_{\underline{\theta}_i}^{\overline{\theta}_i} g(\theta)dF(\theta)}{F(\overline{\theta}_i)-F(\underline{\theta}_i)} & \text{if } \theta \in [\underline{\theta}_i, \overline{\theta}_i). \end{cases}$$

Any element of $\mathrm{MPS}_F(g)$ is an integral over extreme points.

This lemma is a simple generalization of the characterization of extreme points in Kleiner et al. (2021) using a coordinate change. The extreme $P(\theta)$ can be generated by monotone partitional signals, which we define as follows.

Definition 1. *A signal structure (S, σ) is* monotone partitional *if $S = \mathbb{R}$ and $\sigma(\theta) = \delta_{m(\theta)}$ is* degenerate *with $m\colon [\underline{\theta}, \overline{\theta}] \to \mathbb{R}$* monotone.

An extreme point as in Lemma 2 can be implemented by a monotone partitional signal where $m(\theta) = P(\theta)$, that is the signal is revealing when $P(\theta)$ is strictly increasing and pools types that have the same second-order expectation $P(\theta)$.

Novel Proof of Theorem 1 Due to Lemma 1, necessity is easy to show. For sufficiency we use that it is enough to show that extreme points of the monotone functions $P(\theta)$ satisfying (1) are feasible. Then, by Choquet's theorem, all non-extreme points are also feasible by mixing over signal structures (in posterior space).

Necessity: The distribution of $P(\boldsymbol{\theta}) = \mathbb{E}[\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{s}]|\boldsymbol{\theta}]$ is a mean-preserving contraction of the distribution of $\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{s}]$, which is a mean-preserving contraction of the distribution of $\boldsymbol{\theta}$. By Lemma 1, this implies $P \prec_F \mathrm{id}$.

Sufficiency: By Lemma 2, any $P \in \mathrm{MPS}_F(\mathrm{id})$ can be represented as an integral over extreme points. The second-order expectation P is linear under public randomization over signal structures. If the signal spaces of two signal structures are distinct, then randomizing over the signal structures delivers the corresponding mixture over the induced second-order expectations. Thus, if we can

generate the *extreme* $P \in \text{MPS}_F(\text{id})$, we can, after relabeling signal spaces if needed, generate *any* $P \in \text{MPS}_F(\text{id})$ by randomization.[4] The extreme $P\text{MPS}_F(\text{id})$ can be generated by monotone partitional signals, namely by pooling intervals corresponding to the quantile interval where H is constant and fully revealing intervals where H equals the majorizing function. □

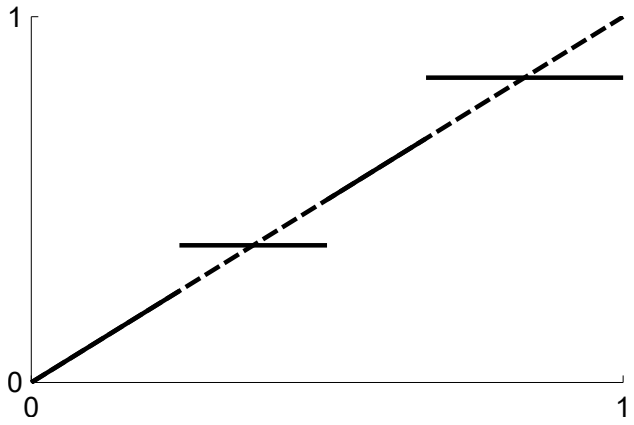


Figure 1: Example of an extreme $P(\theta)$ under $\theta \sim U([0,1])$.

Figure 1 shows an example of an extreme $P(\theta)$. This P can be generated by a monotone partitional signal that reveals each state from 0 to θ_1, pools all states in $[\theta_1, \theta_2)$ together, reveals all states in $[\theta_2, \theta_3)$, and pools all states in $[\theta_3, 1]$.

3.1 Implications

The first corollary follows directly from our proof of Proposition 1, but is worth stating independently.

Corollary 1. *Any feasible monotone $P(\theta) \in L^1(\Theta)$ can be generated by publicly randomizing over monotone partitional signal structures.*

The corollary states that we can generate monotone feasible second-order expectation by using very simple and widely used signal structures, namely monotone partitional signal structures. These signal structures resemble grading or rating schemes with discrete categories, which pool together intervals of the underlying variable of interest. A simple example of a monotone partitional signal structure is a grading scheme where all grades greater or equal to 90 are pooled as A's, all grades greater or equal to 80 but below 90 are pooled as B's, etc.

Second, we can use the connection to majorization and extreme points to characterize the optimal signal structures. The insight of our proof is that the extreme monotone second-order expectations can be generated by monotone partitional signals. This is useful because our principal's problem is a linear optimization problem, in which by Bauer's maximum principal, an extreme point is optimal.

[4] A similar extreme point argument is used in Yang and Zentefis (2024) to characterize the set of feasible distributions of posterior quantiles.

To state our result, let $h(\theta) := \theta - \frac{1-F(\theta)}{f(\theta)}$ denote the virtual valuation of type $\theta \in \Theta$. Further, define $H(\theta) := \int_{\underline{\theta}}^{\theta} h(\theta')d\theta'$ and let $\text{conv}(H)$ be the largest convex function that lies below H.

Corollary 2. *There exists a solution to the principal's problem P, that employs a monotone partitional signal structure.*

More specifically, if the virtual valuation is non-decreasing, then full revelation is optimal. Otherwise, the optimal monotone partitional signal pools exactly those intervals of types on which $\text{conv}(H)$ *is affine.*

Proof. This result follows directly from section 3.2.1 in Kleiner et al. (2021), when transforming the rewritten principal's problem P' into quantile space, combined with the insight that extreme monotone second-order expectations can be generated by monotone partitional signals, which follows from our proof of Theorem 1. □

This result generalizes and reestablishes the results from Rayo (2013) and Onuchic and Ray (2023) through majorization techniques. Both papers study versions of our principal's problem but restrict for tractability to monotone partitional signals.[5] Corollary 2 shows that this restriction is without loss. Further, it reestablishes their characterizations of the optimal signal structures under non-monotonic virtual valuation through majorization techniques.

It is important to note the limits of Corollary 2. Our principal's rewritten problem P' is linear in $P(\theta)$ because we have assumed a quasi-linear environment: the agent's cost and the principal's value are linear in the action. This is a useful benchmark and reasonable in some circumstances (see Rayo, 2013). The more complex problem with convex costs for the agent, in which case a non-extreme point can be optimal, is studied in Smith (2024). Corollary 1 implies nevertheless that we can generate the solution through randomization over monotone partitional signals.

Finally, and beyond the screening problem studied in this paper, our proof of Theorem 1 can be used to characterize higher-order expectations. We can define higher-order expectations $P_n(\theta)$ generated by a signal structure (S, σ), recursively via

$$P_2(\theta) := P(\theta)$$
$$P_n(\theta) := \mathbb{E}[\mathbb{E}[P_{n-1}(\boldsymbol{\theta})|\boldsymbol{s}]|\boldsymbol{\theta} = \theta]$$

Intuitively, the third-order expectation can be understood as follows. Conditional on signal $\boldsymbol{s} = s$, an observer can form an expectation about the second-order expectation of the agent whose type is observed through s. The third-order expectation tells us the expectation that the agent of type θ has about the expectations other types have about her second-order expectation.

Corollary 3. *The set of feasible monotone* $P_n(\theta)$ *coincides with the set of feasible monotone* $P(\theta)$*, namely* $MPS_F(id)$*.*

[5] To be precise, Rayo, 2013 restricts to partitional signal structures but due to the incentive constraint it follows immediately, that we can restrict attention to monotone partitional signals.

Proof. The proof proceeds similar to the proof of Proposition 1. We need to show that any feasible monotone $P_n(\theta)$ is majorized by the identity function with respect to F.

Necessity: By construction, the distribution of $P_n(\boldsymbol{\theta})$ is a mean-preserving contraction of the distribution of $P_{n-1}(\boldsymbol{\theta})$. By induction, for any n, the distribution of $P_n(\boldsymbol{\theta})$ is a mean-preserving contraction of $\boldsymbol{\theta}$. By Lemma 1, this implies $P \prec_F \text{id}$.

Sufficiency: Because a monotone partitional signal is deterministic, all higher-order expectations are identical. More precisely, conditional on a signal realization, the second-order expectation is deterministic and thus equal to the third-order expectation,

$$\mathbb{E}[\underbrace{\mathbb{E}[P(\boldsymbol{\theta})|\boldsymbol{s}]}_{P(\boldsymbol{\theta})}|\boldsymbol{\theta}=\theta] = \mathbb{E}[P(\boldsymbol{\theta})|\boldsymbol{\theta}=\theta] = P(\theta).$$

Inductively, all higher-order expectations are the same. By our proof of Proposition 1, the extreme points of $\text{MPS}_F(\text{id})$ can be generated by publicly randomizing over monotone partitional signals. Higher-order expectations, like second-order expectations, are linear under public randomization, thus we can generate all $P_n(\theta) \in \text{MPS}_F(\text{id})$. □

4 Risk Aversion over Image

This section shows that under risk aversion over image, full privacy can be uniquely optimal. That is, the incentive instrument is *not used* and the agent is not screened. We contrast this with incentive provision through allocating a non-informational good, such as money. In that case, the instrument is *always used* and the principal does screen the agent, in the sense that more of the good is allocated to higher types. To make this comparison in the simplest way, we consider two versions a simple screening problem with *binary types*: first, under incentive provision through image and second, under incentive provision through allocating a non-informational good like money.

To model risk aversion, we assume that the agent values the belief $\pi \in \Delta(\Theta)$ that the population holds by $v(\pi)$ with v concave. For example, the agent's utility may be a concave function of the posterior mean, $v(\pi) = \hat{v}(\mathbb{E}_\pi[\theta])$ with $\hat{v}$ concave. Such risk aversion over image can be stem from agents intrinsic valuation of image due to psychological reasons. Butera et al. (2022) provide evidence for risk aversion over prosocial image, which can be interpreted in such a way. Alternatively, risk aversion over image can be induced from risk aversion over others goods if image is valued for instrumental reasons. For example, consider an agent who values image over her type, say ability, because her attainable wage in the labor market equals her expected ability. If she has risk-averse preferences over income, this induces risk aversion over image, because her utility will be a concave function of her expected ability.

Principal's Problem To illustrate the mechanism in the simplest example, we restrict attention to binary types, $\Theta = \{\underline{\theta}, \overline{\theta}\}$. We identify the set of beliefs $\Delta(\Theta)$ with the unit interval $[0,1]$ and let the prior $\mu \in \Delta(\Theta)$ over types be non-degenerate, that is, $\mu \in (0,1)$. Otherwise there is

effectively only one type and the problem is trivial. We formulate a general version of the principal's problem (P) which allows for the incentive payoffs V to be provided through image or through a non-informational good by appropriately specifying the set $\mathcal{F}$ of feasible payoffs.

$$\max_{a,V\colon \Theta\to\mathbb{R}} (1-\mu)a(\underline{\theta}) + \mu a(\overline{\theta}) \tag{P}$$

s.t.

$$\overline{\theta}V(\overline{\theta}) - a(\overline{\theta}) \geq \overline{\theta}V(\underline{\theta}) - a(\underline{\theta}) \qquad \forall\theta \in \{\underline{\theta},\overline{\theta}\} \tag{IC}$$

$$\underline{\theta}V(\underline{\theta}) - a(\underline{\theta}) \geq \theta\cdot\underline{\theta} \qquad \forall\theta \in \{\underline{\theta},\overline{\theta}\} \tag{IR}$$

$$V \in \mathcal{F} \tag{F}$$

We consider two versions of the feasibility constraint (F), which correspond to incentive provision through image and to incentive provision through a non-informational good, respectively. Before we spell out these two versions, recall that, by standard arguments, the principal's problem can be written as a linear optimization problem over the incentive payoffs V subject to monotonicity and feasibility, that is,

$$\max_{V:\Theta\to\mathbb{R}} (1-\mu)h(\underline{\theta})V(\underline{\theta}) + \mu h(\overline{\theta})V(\overline{\theta}) \tag{P'}$$

s.t.

$$V(\underline{\theta}) \leq V(\overline{\theta}) \tag{MON}$$

$$V \in \mathcal{F} \tag{F}$$

In the Appendix 5.3, we show that, under two types and any prior μ, the virtual valuations $(h(\underline{\theta}), h(\overline{\theta}))$ are strictly increasing, $h(\underline{\theta}) = \underline{\theta} - \frac{\mu}{1-\mu}(\overline{\theta}-\underline{\theta}) < \overline{\theta} = h(\overline{\theta})$.

Incentive Provision through Image First, we consider the feasible incentive payoffs that can be induced by some revealing some information about the agent's type when there are image concerns. As introduced above, we let the value of belief $\pi \in \Delta(\Theta) = [0,1]$ to each agent be $v(\pi)$ where $v\colon [0,1] \to \mathbb{R}$ is strictly increasing and concave. Further, we assume that v is twice differentiable.

Assumption 1. *The value of image $v \in C^2([0,1],\mathbb{R})$ is strictly increasing and concave.*

We say the payoff $V : \Theta \to \mathbb{R}$ is *generated* by a distribution $\tau \in \Delta([0,1])$ over posteriors if $V(\theta)$ corresponds to the expected value $v(\pi)$ of the posterior π, conditioning on the true state being θ, that is,

$$V(\theta) = \mathbb{E}_\tau[v(\pi)|\boldsymbol{\theta} = \theta] \qquad \forall\theta \in \Theta.$$

We let the feasible set of image payoffs $\mathcal{F}_I(v)$ be the set of functions $V\colon \Theta \to \mathbb{R}$ that are generated by some Bayes-consistent distribution $\tau \in \Delta([0,1])$ over posteriors π, that is, τ which satisfy

$\mathbb{E}_\tau[\pi] = \mu$. That means, we allow the principal to choose *any* signal structure without parametric restrictions.

The following theorem shows that when the level of absolute risk aversion over image is pointwise high enough, then full privacy is uniquely optimal.

Theorem 2. *Consider the principal's problem (P) with incentive provision through image, that is,* $\mathcal{F} = \mathcal{F}_I(v)$. *If* $\forall \pi \in [0,1]$:

$$\frac{-v''(\pi)}{v'(\pi)} > \frac{2(h(\overline{\theta}) - h(\underline{\theta}))}{(1-\pi)h(\underline{\theta}) + \pi h(\overline{\theta})}, \tag{2}$$

then no information revelation is uniquely optimal.

We prove this result in the Appendix, where we show that one can rewrite the principal's problem as a standard information problem. We show that the under the provided condition, the value function is strictly concave, so no information revelation is uniquely optimal by Jensen's inequality.

Theorem 2 shows that the incentive instrument, revealing information, is sometimes not used and the agent is not screened. Next, we contrast this result with the standard setup, when incentives are provided through allocating some non-informational good. In that case, we show that the incentive instrument is always used when the agent's utility is differentiable.

Incentive Provision through Allocating a Non-Informational Good Second, we consider the feasible payoffs that can be induced by distributing among the agents some budget of a divisible good like money. We say the payoff $V : \Theta \to \mathbb{R}$ is *induced* by some allocation $x : \Theta \to \mathbb{R}_{\geq 0}$ under utility function $u : \mathbb{R}_{\geq 0} \to \mathbb{R}$ and budget $b \in \mathbb{R}_{>0}$ if

$$\begin{aligned} V(\theta) &= u(x(\theta)) && \forall \theta \in \Theta \\ b &\geq (1-\mu)x(\underline{\theta}) + \mu x(\overline{\theta}). \end{aligned}$$

The first line states that $V(\theta)$ is the utility of obtaining $x(\theta)$ units of the good. The second line states that the allocation x does not exceed the budget b. We let the feasible set of allocational payoffs $\mathcal{F}_{\neg I}(u,b)$ be the set of payoff functions $V\colon \Theta \to \mathbb{R}$ that are induced by some allocation x under utility u and budget b. That is, we allow the principal to choose any budget-balanced allocation. At the expense of some additional terminology, we could also allow for random allocations or allocating multiple goods; this would not change our results.

The following proposition shows that no matter the level of risk aversion over the non-informational good, as long as the utility u is differentiable at b, the agents will be screened.

Proposition 1. *Consider the principal's problem (P) with incentive provision through non-informational goods, that is,* $\mathcal{F} = \mathcal{F}_{\neg I}(u,b)$. *If* u *is differentiable at* b *with* $u'(b) > 0$, *then the optimal mechanism allocates more of the good to the high type,* $x(\overline{\theta}) > x(\underline{\theta})$.

The intuition for this result is as follows. If the utility u is differentiable at b, then u is locally approximately linear at b.[6] Thus, the utility loss from risk aversion is second-order for small deviations from the "equal allocation" $x(\underline{\theta}), x(\overline{\theta})) = (b, b)$, which allocates to both types the same. On the other hand, deviating from the equal allocation by allocating slightly more to the higher type and slightly less to the lower type, provides an *incentive* to the high type to choose a higher action. We show this gain from the generated incentive is first-order and therefore dominates the second-order loss from risk aversion.

Why does a similar logic not show that the principal should screen when providing incentives through image? The reason is that the principal has to resort to *random* allocations of image, that is, random signal structures, to implement small incentives. To explain the difference between image and non-informational goods, we first make precise the trade-off the principal has between the aggregate loss due to risk aversion and creating an incentive for the agent.

Image versus Non-Informational Goods To understand the difference between image and non-informational goods, it is useful to think of the principal as maximizing a linear combination of the *incentive* $I := V(\overline{\theta}) - V(\underline{\theta})$ and the *aggregate payoff* $A := (1-\mu)V(\underline{\theta}) + \mu V(\overline{\theta})$. In our screening problem, as usual, the principal's objective can be written as a linear combination of the allocation of incentive payoffs $V(\theta)$, see (P'). As we show in Appendix 5.3, we can, through a linear coordinate change, express the principal's objective U as a linear combination of the incentive I and the aggregate image utility A, where the weights on both I and A are necessarily positive,

$$U = (1-\mu)h(\underline{\theta})V(\underline{\theta}) + \mu h(\overline{\theta})V(\overline{\theta}) = \underbrace{\mu(1-\mu)(h(\overline{\theta}) - h(\underline{\theta}))}_{>0} I + \underbrace{((1-\mu)h(\underline{\theta}) + \mu h(\overline{\theta}))}_{>0} A. \qquad (3)$$

We prove Proposition 1 for non-informational goods by showing that deviating from the equal allocation and allocating ε more units to the high type provides is a *first-order gain* to the incentive, I, but only a *second-order loss* of aggregate payoff A from risk aversion. By the principal having a finite marginal rate of substitution between incentive I and aggregate payoff A, for small ε the principal gains overall.

By contrast, when it comes to image, deviating from no information revelation, *the incentive gain and the aggregate loss are of the same order.* Thus, if the principal cares enough about the incentive vis-à-vis the aggregate payoff, they choose to reveal no information.

This is illustrated in Figure 2, which plots the Pareto frontier of feasible $(V(\underline{\theta}), V(\overline{\theta}))$ under image and under a non-informational good. For comparison, we chose $v = u$ and $b = \mu$. The bottom right corner of the Pareto frontiers correspond to full privacy and the equal allocation, respectively. Moving from that point along the dotted line preserves the aggregate payoff A but increases the incentive I and thereby improves the principal's objective. Thus, the linear indifference curves of the principal is less steep than the dotted line. This rules out a corner solution for the Pareto frontier for non-informational goods, but it allows for a corner solution for the Pareto frontier for

[6] Footnote on Loss aversion

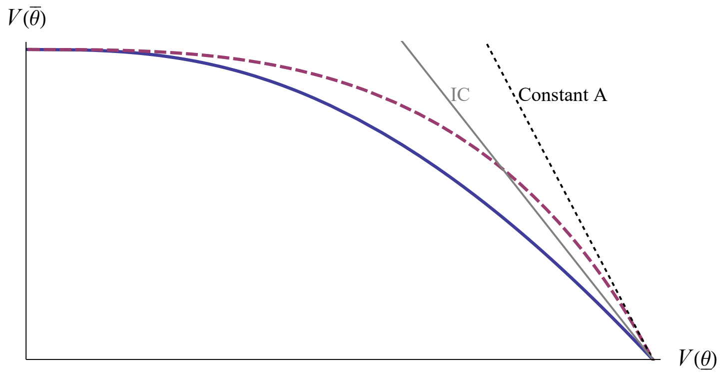
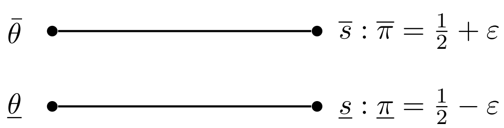


Figure 2: Pareto frontiers under incentive provision through image and a non-informational good.

image, for example, under the gray indifference curve, labeled IC, of the principal.

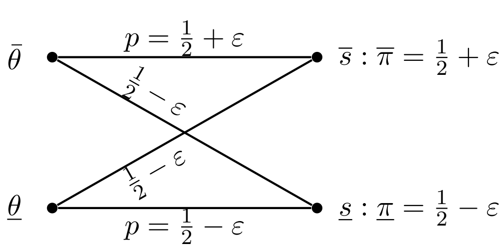
Why are the incentive gain and the aggregate loss of the same order when it comes to image, by contrast to non-informational goods? The reason is that the principal has to *randomize* to achieve a small image incentive, which is not needed when creating an incentive with a divisible non-informational good. The following example illustrates. Suppose, the principal can divide one unit of a divisible non-informational good between two agents. To create a small incentive, they can allocate $1/2+\varepsilon$ units to the high type and $1/2-\varepsilon$ units to the low type, which induces the indicated incentive payoffs $(V(\underline{\theta}), V(\overline{\theta}))$.



$$\Rightarrow V(\overline{\theta}) = u\left(\tfrac{1}{2}+\varepsilon\right)$$

$$\Rightarrow V(\underline{\theta}) = u\left(\tfrac{1}{2}-\varepsilon\right)$$

By contrast, when it comes to image, the principal can only offer a small incentive only by randomizing. That is because the only non-random signal structures either reveal no information (and thereby create no incentive) or are fully revealing (and create the maximal incentive). For example, the following signal structure sends with probability $1/2+\varepsilon$ the high signal $\overline{s}$ conditional on the high type $\overline{\theta}$ and similarly for the low type. The posteriors associated with the high and low signals are $\overline{\pi} = \frac{1}{2}+\varepsilon$ and $\underline{\pi} = \frac{1}{2}-\varepsilon$ if the prior is $\mu = 1/2$. The two types now receive random image payoffs, which induces the indicated incentive payoffs $((V(\underline{\theta}), V(\overline{\theta}))$.



$$\Rightarrow V(\overline{\theta}) = \left(\tfrac{1}{2}+\varepsilon\right) v\left(\tfrac{1}{2}+\varepsilon\right) + \left(\tfrac{1}{2}-\varepsilon\right) v\left(\tfrac{1}{2}-\varepsilon\right)$$

$$\Rightarrow V(\underline{\theta}) = \left(\tfrac{1}{2}+\varepsilon\right) v\left(\tfrac{1}{2}-\varepsilon\right) + \left(\tfrac{1}{2}-\varepsilon\right) v\left(\tfrac{1}{2}+\varepsilon\right)$$

Now, let us compare the aggregate loss and incentive in both cases under $u = v$. In both cases, half the agents obtain utility $u\left(\frac{1}{2}+\varepsilon\right) = v\left(\frac{1}{2}+\varepsilon\right)$ and half the agents obtain utility $u\left(\frac{1}{2}-\varepsilon\right) = v\left(\frac{1}{2}-\varepsilon\right)$ (albeit randomly under image). Thus, the aggregate loss from concavity of v is the same

in both cases, and, in particular, second order in ε. However, the incentive $V(\overline{\theta}) - V(\underline{\theta})$ is smaller under image than under the non-informational good, because the higher type receives the higher belief only some of the time. In particular, one can show that the incentive is first-order in ε under the non-informational good, but second-order in ε under image. Thus, under a non-informational good, the incentive gain dominates the aggregate loss for small enough ε while this is not given under image. Our proof of Theorem 2 implies that, for image, the incentive gain and aggregate loss are of the same order, regardless of which signal structure is used.

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, *54*(2), 442–492.

Albano, G. L., & Lizzeri, A. (2001). Strategic Certification and Provision of Quality. *International Economic Review*, *42*(1), 267–283.

Ali, S. N., & Bénabou, R. (2020). Image versus Information: Changing Societal Norms and Optimal Privacy. *American Economic Journal: Microeconomics*, *12*(3), 116–164.

Alonso, R., & Câmara, O. (2016). Persuading voters. *American Economic Review*, *106*(11), 3590–3605.

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5), 1652–1678.

Bénabou, R., & Tirole, J. (2011). *Laws and norms* (tech. rep.). National Bureau of Economic Research.

Blumkin, T., Margalioth, Y., & Sadka, E. (2015). Welfare stigma re-examined. *Journal of Public Economic Theory*, *17*(6), 874–886.

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*, 131–153.

Butera, L., Metcalfe, R., Morrison, W., & Taubinsky, D. (2022). Measuring the welfare effects of shame and pride. *American Economic Review*, *112*(1), 122–68.

Daughety, A. F., & Reinganum, J. F. (2010). Public Goods, Social Pressure, and the Choice between Privacy and Publicity. *American Economic Journal: Microeconomics*, *2*(2), 191–221.

Dewatripont, M., Jewitt, I., & Tirole, J. (1999). The Economics of Career Concerns, Part I: Comparing Information Structures. *The Review of Economic Studies*, *66*(1), 183–198.

Doval, L., & Smolin, A. (2024). Persuasion and Welfare. *Journal of Political Economy*, *132*(7), 2451–2487.

Dubey, P. K., & Geanakoplos, J. (2014, July). Games with Money and Status: How Best to Incentivize Work.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1934). *Inequalities*. Cambridge University Press.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The review of Economic studies*, *66*(1), 169–182.

Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615.

Kleiner, A., Moldovanu, B., & Strack, P. (2021). Extreme points and majorization: Economic applications. *Econometrica*, *89*(4), 1557–1593.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The review of economic studies*, *38*(2), 175–208.

Onuchic, P., & Ray, D. (2023). Conveying value via categories. *Theoretical Economics*, *18*(4), 1407–1439.

Posner, R. A. (1978). The Right of Privacy. *Georgie Law Review*, *12*, 393–422.

Rayo, L. (2013). Monopolistic Signal Provision. *The B.E. Journal of Theoretical Economics*, *13*(1), 27–58.

Saeedi, M., & Shourideh, A. (2022). Optimal Rating Design.

Saeedi, M., & Shourideh, A. (2024). Optimal Rating Design under Moral Hazard.

Shaked, M., & Shanthikumar, J. G. (Eds.). (2007). *Stochastic Orders*. Springer New York.

Smith, A. (2024). Optimal Signalling Incentives. *Working Paper*.

Tirole, J. (2021). Digital dystopia. *American Economic Review*, *111*(6), 2007–2048.

Zapechelnyuk, A. (2020). Optimal Quality Certification. *American Economic Review: Insights*, *2*(2), 161–176.

Zubrickas, R. (2015). Optimal Grading. *International Economic Review*, *56*(3), 751–776.

5 Appendix

Additional assumption: F has continuous positive density.

5.1 Proof of Lemma 1

Proof. It is well known that the distribution of $P(\boldsymbol{\theta})$ is a mean-preserving contraction of the distribution of $\boldsymbol{\theta}$ if and only if the CDF G of $P(\boldsymbol{\theta})$ majorizes the CDF F of $\boldsymbol{\theta}$, that is,

$$\forall \hat{\theta} \in [\underline{\theta}, \overline{\theta}]\colon \int_{\hat{\theta}}^{\overline{\theta}} F(\theta)d\theta \leq \int_{\hat{\theta}}^{\overline{\theta}} G(\theta)d\theta,$$

with equality at $\hat{\theta} = \underline{\theta}$. The CDF G majorizes F if and only if F^{-1} majorizes G^{-1} (e.g., Shaked and Shanthikumar, 2007, Theorem 3.A.5.), that is,[7]

$$\forall \hat{q} \in [0,1]\colon \int_{\hat{q}}^{1} G^{-1}(q)dq \leq \int_{\hat{q}}^{1} F^{-1}(q)dq,$$

[7]For non-continuous monotonic F, F^{-1} denotes the right-continuous inverse $F^{-1}(q) := \sup\{x|F(x) \leq q\}$ for $q \in [0,1]$.

with equality at $\hat{q} = 0$. Under comonotonicity (so for all θ, $P(\theta)$ and θ correspond to the same quantile) this is equivalent to

$$\forall \hat{q} \in [0,1]\colon \int_{\hat{q}}^{1} P(F^{-1}(q))dq \leq \int_{\hat{q}}^{1} F^{-1}(q)dq, \tag{4}$$

with equality at $q = 0$. Recall that by assumption, F is continuously differentiable and strictly increasing. Thus, we can apply a coordinate change to $q = F(\theta)$ and rewrite the (4) as

$$\forall \hat{\theta} \in [\underline{\theta}, \overline{\theta}]\colon \int_{\hat{\theta}}^{\overline{\theta}} P(\theta)dF(\theta) \leq \int_{\hat{\theta}}^{\overline{\theta}} \theta dF(\theta).$$

□

5.2 Proof of Lemma 2

Proof. The proof carries over the Kleiner et al. (2021) characterization of extreme points by going into quantile space.

Let

$$\mathrm{MPS}_F(g) := \{P \in L^1(\mathbb{R}) | P \text{ is monotone and } P \prec_F g\}.$$

That is, $\mathrm{MPS}_F(g)$ contains the functions $P \in L^1(\mathbb{R})$ such that

$$\begin{aligned} \int_{\hat{\theta}}^{\overline{\theta}} P(\theta)dF(\theta) &\leq \int_{\hat{\theta}}^{\overline{\theta}} g(\theta)dF(\theta) \quad \forall \hat{\theta} \in (\underline{\theta}, \overline{\theta}], \\ \int_{\underline{\theta}}^{\overline{\theta}} P(\theta)dF(\theta) &= \int_{\underline{\theta}}^{\overline{\theta}} g(\theta)dF(\theta). \end{aligned}$$

Going via the coordinate change $q = F(\theta)$ into quantile space, this is equivalent to[8]

$$\begin{aligned} \int_{\hat{q}}^{1} H(q)dq &\leq \int_{\hat{q}}^{1} g(F^{-1}(q))dq \quad \forall \hat{q} \in (0,1] \\ \int_{0}^{1} H(q)dq &= \int_{0}^{1} g(F^{-1}(q))dq \end{aligned} \tag{5}$$

where $H(q) := P(F^{-1}(q))$. Let

$$\mathrm{MPS}(g \circ F^{-1}) := \{H \in L^1([0,1]) | H \text{ is monotone and satisfies (5)}\}.$$

The characterization of extreme points from Kleiner et al., 2021, Theorem 1, applies to $\mathrm{MPS}(g \circ F^{-1})$.

[8] Recall that F^{-1} denotes the right-continuous inverse $F^{-1}(q) := \sup\{x | F(x) \leq q\}$ for $q \in [0,1]$. Thus, we have $F^{-1}(F(\theta)) = \sup\{\theta' | F(\theta') \leq F(\theta)\} = \theta$ by right-continuity of F, so $P(\theta) = P(F^{-1}(F(\theta)))$.

We use the following function to show that the characterization applies analogously to $\text{MPS}_F(g)$:

$$\begin{aligned}\gamma\colon \text{MPS}_F(g) &\to \text{MPS}(g(F^{-1}))\\ P &\mapsto H = P \circ F^{-1}\end{aligned}$$

From the above, one can easily verify that γ is well-defined, linear, and preserves the L^1-norm.

By assumption, F is strictly increasing and continuous, thus $F : [\underline{\theta}, \overline{\theta}] \to [0, 1]$ is invertible. Then, γ has an inverse $\gamma'\colon H \mapsto P = H \circ F$, so γ is a bijection. Thus, γ is an isomorphism on normed vector spaces and both the Choquet representation (Proposition 1) and the characterization of extreme points (Theorem 1) from Kleiner et al. (2021) immediately carry over to $\text{MPS}_F(g)$.

□

5.3 Screening with Two Types

By standard arguments, the IR constraint holds for low type and the IC constraint holds for the high type, so we get the following maximization problem.

$$\begin{aligned}&\max(1-\mu)a(\underline{\theta}) + \mu a(\overline{\theta})\\ &\qquad s.t.\\ &\overline{\theta}V(\overline{\theta}) - a(\overline{\theta}) = \overline{\theta}V(\underline{\theta}) - a(\underline{\theta}) && (\text{IC}(\overline{\theta}))\\ &\underline{\theta}V(\underline{\theta}) - a(\underline{\theta}) = \underline{\theta}^2 && (\text{IR}(\underline{\theta}))\end{aligned}$$

We can rewrite the objective as follows:

$$\begin{aligned}a(\underline{\theta}) &= \underline{\theta}V(\underline{\theta}) - \underline{\theta}^2\\ a(\overline{\theta}) &= a(\underline{\theta}) + \overline{\theta}(V(\overline{\theta}) - V(\underline{\theta})) = \overline{\theta}V(\overline{\theta}) - (\overline{\theta} - \underline{\theta})V(\underline{\theta}) - \underline{\theta}^2\\ \Rightarrow (1-\mu)a(\underline{\theta}) + \mu a(\overline{\theta}) &= V(\underline{\theta})((1-\mu)\underline{\theta} - \mu(\overline{\theta} - \underline{\theta})) + \mu\overline{\theta}V(\overline{\theta}) - \underline{\theta}^2\end{aligned}$$

Modulo a constant, the objective is

$$(1-\mu)h(\underline{\theta})V(\underline{\theta}) + \mu h(\overline{\theta})V(\overline{\theta})$$

where

$$\begin{aligned}h(\underline{\theta}) &:= \underline{\theta} - \frac{\mu}{1-\mu}\left(\overline{\theta} - \underline{\theta}\right) < \underline{\theta}\\ h(\overline{\theta}) &:= \overline{\theta}\end{aligned}$$

and thus $h(\underline{\theta}) < \underline{\theta} < \overline{\theta} = h(\overline{\theta})$, or $h(\overline{\theta}) - h(\underline{\theta}) > 0$.

Finally, we rewrite our objective as a linear combination of $A = \mu V(\underline{\theta}) + (1-\mu)V(\overline{\theta})$ and

$I = V(\overline{\theta}) - V(\underline{\theta})$, we get

$$V = \underbrace{\big((1-\mu)h(\underline{\theta}) + \mu h(\overline{\theta})\big)}_{=\underline{\theta}>0} A + \mu(1-\mu)\underbrace{\big(h(\overline{\theta}) - h(\underline{\theta})\big)}_{>0} I.$$

That is, the principal can be understood as caring about a linear combination of aggregate image utility A and incentive I with positive weights each.

5.4 Proof of Theorem 2

Proof. We prove this result by reformulating the principal's problem as a standard information principal's problem with a strictly concave value function, making no information revelation uniquely optimal.

Let $\Delta^{\mu}([0,1]) \subset \Delta([0,1])$ denote the set of distributions τ over posteriors π that are Bayes-consistent, that is, $\mathbb{E}_{\tau}[\pi] = \mu$. We adopt the posterior approach and write the principal's objective U_{τ} as a function of the Bayes-consistent distribution $\tau \in \Delta^{\mu}([0,1])$ over posteriors $\pi \in [0,1]$,

$$U_{\tau} = (1-\mu)h(\underline{\theta})V_{\tau}(\underline{\theta}) + \mu h(\overline{\theta})V_{\tau}(\overline{\theta}), \tag{6}$$

where $V_{\tau}(\theta) = \mathbb{E}_{\tau}[v(\pi)|\boldsymbol{\theta} = \theta]$ is the expected value of the posterior conditional on the true state being θ. Ignoring the monotonicity constraint, $V_{\tau}(\underline{\theta}) \leq V_{\tau}(\overline{\theta})$ for now, the principal's problem is simply choosing $\tau \in \Delta^{\mu}([0,1])$ to maximize U_{τ}. Conveniently, one can reformulate this conditional expectation as an unconditional expectation with respect to τ (see also Alonso and Câmara, 2016; Doval and Smolin, 2024):

$$V_{\tau}(\underline{\theta}) := \mathbb{E}_{\tau}\Big[\frac{1-\pi}{1-\mu}v(\pi)\Big], \tag{7}$$

$$V_{\tau}(\overline{\theta}) := \mathbb{E}_{\tau}\Big[\frac{\pi}{\mu}v(\pi)\Big]. \tag{8}$$

Inserting (7) and (8) into (6), we rewrite the objective as an unconditional expectation as in standard information principal's problems,

$$U_{\tau} = \mathbb{E}_{\tau}[\underbrace{\big((1-\pi)h(\underline{\theta}) + \pi h(\overline{\theta})\big)v(\pi)}_{\tilde{v}(\pi):=}].$$

Taking the derivative of the value function $\tilde{v}(\pi)$ with respect to $\pi \in [0,1]$ gives

$$\tilde{v}'(\pi) = (h(\overline{\theta}) - h(\underline{\theta}))v(\pi) + ((1-\pi)h(\underline{\theta}) + \pi h(\overline{\theta}))v'(\pi)$$
$$\tilde{v}''(\pi) = 2(h(\overline{\theta}) - h(\underline{\theta}))v'(\pi) + ((1-\pi)h(\underline{\theta}) + \pi h(\overline{\theta}))v''(\pi)$$

Under $v'(\pi) \neq 0$, we have

$$\tilde{v}''(\pi) < 0 \Leftrightarrow \frac{-v''(\pi)}{v'(\pi)} > \frac{2(h(\overline{\theta}) - h(\underline{\theta}))}{(1-\pi)h(\underline{\theta}) + \pi h(\overline{\theta})}.$$

Let $\delta(\mu) \in \Delta([0,1])$ denote the degenerate distribution that reveals no information. Under a strictly concave value function $\tilde{v}(\pi)$, Jensen's inequality implies

$$U_\tau = \mathbb{E}_\tau[\tilde{v}(\pi)] \leq \tilde{v}(\mathbb{E}_\tau[\pi]) = \tilde{v}(\mu) = U_{\delta(\mu)}$$

with a strict inequality if τ is not degenerate, $\tau \neq \delta(\mu)$. Because $\tau = \delta(\mu)$ satisfies the monotonicity constraint, $V_\tau(\underline{\theta}) \leq V_\tau(\overline{\theta})$, no information revelation is uniquely optimal. □

5.5 Proof of Proposition 1

Proof. Consider the allocation $(b - \frac{\mu}{1-\mu}\varepsilon, b+\varepsilon)$ for $\frac{1-\mu}{\mu}b \geq \varepsilon > 0$. This allocation is budget balanced for all ε. We show that for ε small enough, this allocation satisfies the monotonicity constraint and improves upon the equal allocation (b, b).

Recall that the principal's objective is a linear combination of the aggregate image utility $A = \mu V(\underline{\theta}) + (1-\mu)V(\overline{\theta})$ and incentive $I = V(\overline{\theta}) - V(\underline{\theta})$. By u being differentiable at b, we can apply Taylor's theorem and rewrite A and I as functions of ε up to second-order terms in ε as:

$$A(\varepsilon) = (1-\mu)u(b - \frac{\mu}{1-\mu}\varepsilon) + \mu \cdot u(b+\varepsilon) = u(b) + \underbrace{\left(\mu\varepsilon - (1-\mu)\frac{\mu}{1-\mu\varepsilon}\right)}_{=0} u'(b) + o(\varepsilon^2)$$

$$I(\varepsilon) = u(b+\varepsilon) - u(b - \frac{\mu}{1-\mu}\varepsilon) = \frac{1}{1-\mu}u'(b)\varepsilon + o(\varepsilon^2)$$

Thus, under $u'(b) > 0$ there is a first-order gain to the incentive I, whereas the change to the aggregate image utility A is second-order. Hence, even if the change to the aggregate image utility A is negative (which it is under concave u), the positive effect on the incentive dominates for small ε as long as the principal has a finite marginal rate of substitution between A and I. Let the principal's weights on A and I be h_A and h_I, respectively. By section 5.3, both weights are positive, $h_A, h_I > 0$. Taking the difference between the achieved value under $\varepsilon > 0$ and the value under $\varepsilon = 0$, which corresponds to the equal allocation, we get

$$\lim_{\varepsilon \to 0} (h_A A(\varepsilon) + h_I I(\varepsilon)) - h_A v(b) = \lim_{\varepsilon \to 0} \underbrace{h_I \frac{1}{1-\mu} v'(\mu)}_{>0} \varepsilon + o(\varepsilon^2) > 0$$

Thus, for ε small enough, the principal's value is necessarily higher than under the equal allocation. Moreover, for such ε, we have $u(b+\varepsilon) > u(b - \frac{\mu}{1-\mu}\varepsilon)$, so the monotonicity constraint is satisfied. □