Regulating Discretion: Information Design with Privately Known Common Interests

Qiyuan Zheng

October 30, 2026

LINK TO MOST RECENT VERSION

Abstract

A principal regulates the information of an agent who privately knows which statistic of the state determines both parties' payoffs from his action. The agent's payoff is this shared component plus a bias, a second statistic of the state. When the statistics are correlated, information that helps the agent learn about the shared statistic may also serve his bias. Our main result characterizes the optimal mechanism, which exhibits two distortions that arise from screening. First, the agent evaluates the state by a virtual statistic rather than his own. Second, the agent conditions on the bias statistic when making his decision. We benchmark the optimal mechanism against simpler alternatives, such as full information and mechanisms that do not condition on the bias statistic. In particular, when the agent can exploit the correlation, forbidding the use of a protected characteristic distorts his evaluation of the state more than the optimal mechanism does.

1 Introduction

Organizations often delegate decisions to agents who are better informed about which criteria to apply. At the same time, they restrict what these agents may learn, as the agents' objectives need not coincide with the organization's. A board of education lets each school admit by its own criterion but regulates whether the admissions process may condition on an applicant's need for financial aid. Similarly, a firm may let each branch make personnel decisions according to its own specialty but regulates whether those decisions can condition on a worker's incumbency. In both instances, the organization shares the agent's criterion when evaluating candidates but wants to curb the agent's bias toward a characteristic of the candidate that it deems irrelevant.

When the organization does not know the agent's criterion, a screening problem arises. The agent can exploit correlation between the attributes that the shared criterion weighs and the characteristic the organization has blinded. By shifting weight towards attributes that reveal the characteristic, he can serve his bias while appearing to apply his criterion. Suppose schools evaluate applicants on grades and extracurriculars, each school with its own weights. Schools also have financial incentives to admit fewer students who need aid. If extracurriculars are more negatively correlated with the need for aid than grades are, a school can circumvent a need-blind policy by overweighting extracurriculars in their admissions process. The same holds for personnel decisions. Branches evaluate workers on technical and client-facing skills, each branch with its own weights. Managers prefer to promote incumbents. If client-facing skill is more positively correlated with incumbency than technical skill is, a manager can circumvent a policy that bars conditioning on incumbency by overweighting client-facing skill. In both cases, the agent has an incentive to misreport his criterion to serve his bias.

Evidence on blinding policies shows the same mechanism at work: the excluded characteristic often re-enters the decision making process through correlated variables. In hiring, ban-the-box laws bar employers from asking about criminal records on job applications. Agan and Starr (2018) show that the policy led employers to use race as a proxy for a criminal record instead. The callback advantage of white applicants over comparable black applicants rose from 7 to 43 percent after ban-the-box laws were implemented. In admissions, Proposition 209 barred the University of California from using race. To recover the drop in minority admission rates, its campuses shifted weight from SAT scores to high school grades and family background, the criteria on which minority applicants compare more favorably (Antonovics and Backes, 2014). These studies show that blinding policies are ineffective and can backfire. A natural question, then, is how an organization should regulate what its agents learn when it disagrees with them about how much a characteristic should matter.

This paper studies an information design model in which the agent's private information

determines how the state enters both his and the principal's payoffs. Each candidate is described by a state (x_1, x_2, θ). The first two dimensions capture the candidate's match attributes. The third dimension θ represents the agent's private benefits from accepting the candidate. In the admissions context, x_1 and x_2 are the applicant's grades and extracurriculars, and θ is the applicant's need for financial aid. The agent's private information, his type τ, is the criterion by which the candidate's match attributes combine into match quality, $x^{(\tau)} \coloneqq \tau x_1 + (1-\tau)x_2$. While both the principal and the agent evaluate a candidate by the same match quality, the agent is more willing to accept candidates that give him higher private benefits. Since the principal does not know the agent's type, she offers a menu of acceptance policies, each of which provides the agent with information about the candidate. The agent observes his type and chooses the policy best suited to his criterion.

Consider first the principal's preferred policy for each type: evaluate the candidate on match quality alone, without regard to θ. We call this the first best policy. A menu of first-best policies is analogous to a blinding regulation: every candidate is evaluated using some agent's criterion with no conditioning on θ. Proposition 2 shows that this menu is incentive compatible if and only if every policy on it delivers the same expected private benefits to the agent. Otherwise some agent gains by switching to the policy of a nearby type that delivers higher private benefits, at little cost in match quality. This result reflects what the studies on blinding policies have shown: the characteristic the organization tries to blind still enters the agent's decision making process through its correlation with attributes the agent is allowed to use.

Our main result characterizes the optimal menu and shows that it generally features acceptance policies that explicitly condition on θ. We first show that every acceptance policy in the optimal menu is a *linear scoring rule*: it scores each candidate by a weighted sum of the match attributes, $\sigma x_1 + (1-\sigma)x_2$, and accepts the candidate if and only if the score exceeds a θ-dependent threshold. Given a menu of linear scoring rules, an agent of type τ picks the rule that yields him the highest payoff, accounting for both his own criterion and the private benefits the rule delivers. Theorem 1 shows that the optimal menu generally exhibits two forms of distortion from the first-best policies. First, the score does not weight the match attributes by the agent's own criterion but by a *virtual criterion* $\sigma \neq \tau$. Second, the threshold depends on θ: a candidate is judged not by their match attributes alone but also by the private benefits they bring to the agent. The theorem also characterizes the two distortions with a first-order condition that captures the following trade-off. The principal can rotate a type's rule toward the shared criterion to raise her payoff from that type while keeping the agent's payoff constant. The same rotation, however, makes that type's rule more attractive to the other types, and their rents must rise to preserve incentive compatibility. At the optimum the two effects balance, and the virtual criterion departs from the agent's own by exactly the cost of those rents.

We then apply our main characterization to settings with ordered private benefits, in which every agent type can obtain a higher expected private benefit by overweighting the same match attribute, call it the favored attribute. A regulation that bars conditioning on θ then fails in a predictable way. Each agent places more weight on the favored attribute than his own preferences over match attributes warrant, as doing so raises his private benefit. Theorem 2 shows that the optimal menu corrects this by explicitly conditioning on θ to compress the spread of private benefits across types and reduce the distortion over the weighting of match attributes. The distortions that remain are one-sided: every agent's weighting still leans toward the favored attribute, and the private benefit delivered along the menu stays ordered across types.

Since the principal's screening problem arises from the correlation between the match attributes and θ, it is natural to ask how her payoff changes when the attributes are less informative about θ. Theorem 3 provides an explicit construction. As match attributes become less informative about θ, any incentive compatible menu can be converted into one that remains incentive compatible and departs from the principal's preferred decision less often. As a consequence, the principal's payoff increases when match attributes become less informative about θ (Corollary 3.1). The key takeaway is that the cost of screening is governed by how much the match attributes reveal about θ.

To solve the problem, we reformulate it as a monopoly screening problem in the spirit of Mussa and Rosen (1978). Identifying each acceptance policy with the surplus it generates on each match attribute and the private benefits it delivers, we characterize incentive compatibility by the standard monotonicity and envelope conditions. The reformulated problem departs from standard screening in two ways. First, every allocation-transfer pair can be delivered by a continuum of acceptance policies. For any allocation-transfer schedule, the principal can therefore select, for each type, the policy that gives her the highest payoff without affecting the agent's incentive constraints. This inner-layer problem already shows that every acceptance policy in the optimal menu is a linear scoring rule. Second, the reformulated problem features endogenous rent terms. More specifically, the principal's cost of conceding rent to the agent depends on how his threshold responds to θ, which is itself part of the solution. In Mussa and Rosen (1978), a unit of rent costs one unit of transfer, so the virtual type is known before the problem is solved; here the virtual criterion is determined along with the menu. The reformulation nonetheless allows us to treat the problem as a control problem over convex rent schedules and to characterize its solution by a first-order condition. Mapping that solution back to acceptance policies through the inner-layer problem delivers our main result: the first-order condition on the rent schedule becomes the balance equation of Theorem 1.

The rest of this section discusses related literature (Section 1.1) and illustrates the mechanism in the admissions setting (Section 1.2). Section 2 sets up the general model. Section 3 characterizes the optimal menu and outlines the analysis behind the solution. Section 4 applies the characterization

to show how the private benefits of the first-best policies shape the optimal menu, and derives comparative statics in the dependence between the match attributes and θ. Section 5 discusses extensions, and Section 6 concludes. Proofs are deferred to a separate appendix

1.1 Related Literature

This paper connects two literatures: the work on the effects of blinding policies, and the theory of information design for privately informed receivers.

The first literature documents that blinding a characteristic does not remove it from decisions. Agan and Starr (2018) find that banning the box raised the racial gap in callbacks at affected employers, and Antonovics and Backes (2014) find that the University of California re-weighted test scores, grades and family background after Proposition 209. The theory of this response goes back to Chan and Eyster (2003), who show that a ban on using race leads admissions offices to re-weight the criteria they may use, and to Fryer et al. (2008), who show that color-blind affirmative action is an inefficient way to reach a given level of representation. Ellison and Pathak (2021) quantify the inefficiency of race-neutral proxies in Chicago's exam schools. Strack and Yang (2024) characterize the signals that reveal nothing about a protected characteristic and the cost of restricting to them. That literature studies what the regulated party does under blindness and what blindness costs. We take the failure of blindness as the starting point and ask what the regulator should design instead, when the party being regulated knows its own criterion and the regulator does not.

The second literature studies a designer who controls the information of a receiver with private information. Kolotilin et al. (2017) show that when the receiver's type enters his payoff as an additive shift of a one-dimensional state, a menu of experiments does no better than a single experiment; Kolotilin (2018) gives the linear-programming treatment of the same environment, and Guo and Shmaya (2019) characterize the optimal single experiment. Candogan and Strack (2023) is the closest paper: receivers with private types and utilities linear in a one-dimensional state, for whom the optimal mechanism is a menu of laminar partitional signals whose complexity grows with the number of types. In our model the type is the weight the agent places on two match attributes, so types rank states differently rather than by a common statistic. This is what gives menus their value and what makes the optimal menu a continuum of linear scoring rules indexed by the type, each distorted in two ways. A second difference is the object being screened. In those papers the designer and the receiver disagree about the state; here they agree about match quality and disagree only about a characteristic that the principal does not value, and the optimal policies condition on that characteristic.

Menus of tests without transfers appear in Li and Xu (2024), where a principal designs tests

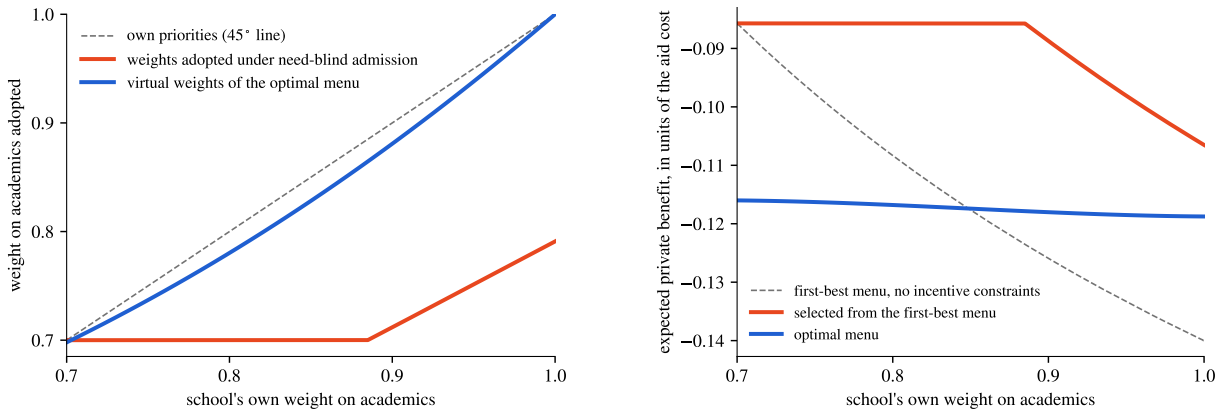
about a one-dimensional state for an agent with a private preference type and the optimal menu has at most three tests, and in Hancart (2025), where the agent's choice from a menu of tests reveals information about the state. Bergemann et al. (2026) study a seller who controls both the buyer's information about her value and the selling mechanism, and find that information rents compress the menu to finitely many items. Transfers change the economics: in our setting information is the only allocation and the only currency, and the rent a type earns is paid in private benefit. On the method, we reformulate the problem as a monopoly screening problem in the spirit of Mussa and Rosen (1978) and Myerson (1981), with the envelope and monotonicity conditions of Milgrom and Segal (2002). The participation constraint binds in the interior of the type space, which places the problem in the class with countervailing incentives studied by Lewis and Sappington (1989), Maggi and Rodríguez-Clare (1995) and Jullien (2000). The absence of transfers connects the problem to delegation, in the tradition of Amador and Bagwell (2013); Kolotilin and Zapechelnyuk (2025) establish an equivalence between persuasion and delegation problems in linear environments, which ours falls outside because the agent's type rotates his ranking of states.

1.2 An Illustration: Admissions with Financial Aid

To fix ideas, consider in more detail the admissions example introduced earlier. A board of education regulates the admission protocols of the schools under its authority. Applicants are described by grades, extracurriculars, and aid status. Each school evaluates an applicant's match quality as a weighted sum of grades and extracurriculars. One can interpret a school's weighting of grades and extracurriculars as its admissions priorities. The board agrees with each school's priorities: it wants a school that prioritizes academics to admit strong academic applicants, and a school that prioritizes extracurriculars to admit strong extracurricular applicants. The school, however, also bears the cost of the aid its admitted students need, a cost the board does not bear. Suppose that applicants who need aid are, on average, as strong on grades as other applicants but weaker on extracurriculars.[1] The natural regulation is need-blind admission: each school admits by its own priorities and ignores aid status.

Figure 1 shows the outcome of need-blind admission when schools choose freely among the protocols it allows. In panel (a), the dashed line is a school's own priorities and the orange line is the weighting of the protocol it actually adopts. Every school adopts weights that favor extracurriculars more than its own priorities do, and the majority of schools adopt the most leadership-heavy protocol

[1] In the example below, 40 percent of applicants need aid; grades are uniformly distributed and unrelated to aid status; extracurriculars of aid applicants are concentrated at low values and those of other applicants at high values; and schools' weights on academics are spread evenly between 0.7 and 1.



(a) Weights

(b) Expected private benefits

Figure 1: Admission Protocols and Benefits. Panel (a): the weight on academics: priorities vs adopted. Panel (b): the expected private benefit

available. The distortion is intuitive: protocols that weight extracurriculars more heavily admit fewer students who need financial aid, and a school that shifts its weights toward extracurriculars loses little match quality while saving aid costs on the students it admits. Under need-blind admission, schools therefore use extracurriculars as a proxy for aid status to serve their financial incentives. Panel (b) reinforces this intuition. The dashed line plots the expected private benefit a school receives from the protocol matching its own priorities, and the orange line the private benefit it receives from the protocol it actually adopts. The gap between the two is the aid cost a school saves by distorting its priorities.

Figure 1 also shows the solution to the board's problem. In the optimal menu, admission protocols explicitly condition on θ. The blue line in panel (b) shows the result: the expected private benefit is nearly the same across the whole menu. With private benefits nearly equal, schools now have reduced incentives to distort their priorities over match attributes. This is also reflected by the blue line in panel (a): the school's weighting over academics and extracurriculars is now much more aligned with its priorities under the optimal menu than it was under the need-blind regulation. The rest of the paper studies the general model and makes these intuitions precise.

2 Framework

A principal (she) regulates the information an agent (he) can use in a binary decision problem. The agent has private information, his *type*, which determines the statistic of the *state* that enters both parties' payoffs. Formally, the state space $\Omega = X_1 \times X_2 \times \Theta$ has three components, $X_i = [0, 1]$

and $\Theta \subseteq \mathbb{R}$. The first two dimensions, (x_1, x_2), are the match attributes of the state. The third dimension, θ, is the agent's *private benefit*: a component of the state that the agent values but the principal does not. In our leading examples, the state describes a candidate. The match attributes are the dimensions on which both parties evaluate the candidate. The private benefit is a characteristic of the candidate the principal deems irrelevant.

The agent's type $\tau \in \mathcal{T} = [0, 1]$ determines the *match quality*

$$x^{(\tau)} := \tau x_1 + (1 - \tau) x_2,$$

a weighted average of the two match attributes. The agent takes action $a = 0, 1$, with payoffs depending on the state ω, the type τ and the action a:

$$\begin{aligned} \text{Principal:} \qquad & v_\tau(\omega, a) = a\left(x^{(\tau)} - \kappa\right), \\ \text{Agent:} \qquad & u_\tau(\omega, a) = a\left(x^{(\tau)} - \kappa + \theta\right), \end{aligned}$$

where $\kappa \in (0, 1)$ is a constant that captures the opportunity cost of taking action $a = 1$. Interpreting the payoffs, while both parties would like the agent to take the action when $x^{(\tau)}$ is high, the agent's decision under perfect information will be biased by θ.

The state is drawn from a common prior $\pi_0 \in \Delta(\Omega)$. The agent's type is drawn independently and uniformly over $\mathcal{T} = [0, 1]$.[2] We impose the following assumptions on the prior over Ω:

Assumption 1.

(A1) $\Theta \subset \mathbb{R}$ is a compact interval with $0 \in \Theta$.

(A2) π_0 admits a density f on Ω that is bounded above, $f \leq f_{\max} < \infty$, and bounded away from zero near the pivotal state $(\kappa, \kappa, 0)$: $f \geq f_{\min} > 0$ on $B\big((\kappa, \kappa, 0), r_0\big) \cap \Omega$ for some $r_0 > 0$.

(A3) $\mathbb{E}[x_i + \theta] < \kappa$ for $i = 1, 2$.

(A1)–(A2) are regularity conditions. (A3) ensures that an uninformed agent will choose $a = 0$, regardless of his type: his uninformed payoff from taking $a = 1$ is a convex combination of the two values in (A3).

Remark. Our results extend to a model where both parties' payoffs depend linearly on θ:

$$\begin{aligned} \text{Principal:} \qquad & v_\tau(\omega, a) = a\left(x^{(\tau)} - \kappa + \beta_P \theta\right), \\ \text{Agent:} \qquad & u_\tau(\omega, a) = a\left(x^{(\tau)} - \kappa + \beta_A \theta\right). \end{aligned}$$

[2]The reduction to direct menus and the incentive analysis of Section 3 hold for any continuous, strictly positive type density. Uniformity is used in the characterization of the optimal menu in the main text, as it rules out bunching and simplifies the main result. Section 5 discusses general type densities.

The crucial assumption for us is that the difference between the principal's and the agent's (state dependent) payoffs does not change with the agent's type.[3]

Example 1 (Admissions with financial aid)**.** A board of education (the principal) regulates the admission protocols of the schools under its authority (the agents). Each applicant is characterized by their academic achievement x_1, extracurricular achievement x_2, and need for financial aid $\theta \leq 0$, with more negative θ's representing greater need. Schools weigh the two achievements differently, and the board shares each school's weighting: it wants a school that prioritizes academics to admit strong academic students, and a school that prioritizes extracurriculars to admit strong extracurricular students. The school, however, also bears the cost of the aid its admitted students need, which the board does not. The board's aim is thus meritocratic admission, with merit measured by the school's own weighting. The school's aim is the same, except it also minds the cost of aid.

Example 2 (Headquarters and branches)**.** A firm's headquarters (the principal) regulates the personnel decisions of its branches (the agents). The state describes a worker: x_1 is technical skill, x_2 is client-facing skill, and θ is the worker's incumbency, which the branch manager values and headquarters does not. A branch's type is the weight its specialty places on technical skill. Client-facing skill is acquired through experience, so it is positively correlated with incumbency. Headquarters' aim is to give branch managers the flexibility to promote the workers best suited to their specialties while limiting the favor shown to incumbents.

The principal regulates the information the agent uses by choosing what he learns about the state. An *acceptance policy* is a measurable map $q : \Omega \to [0, 1]$, where $q(\omega)$ is the probability that action $a = 1$ is recommended in state ω; in the education example, an acceptance policy is an admissions protocol. The principal and agent's payoffs from implementing an acceptance policy are given by

$$U_\tau(q) = \mathbb{E}\big[(x^{(\tau)} - \kappa + \theta)\, q(\omega)\big], \qquad V_\tau(q) = \mathbb{E}\big[(x^{(\tau)} - \kappa)\, q(\omega)\big].$$

A *menu* is a set of acceptance policies. The agent can select a policy from the menu and implement it, or opts out and chooses an uninformed $a = 0$ action. Two modelling commitments are built in: the principal commits to the menu, and an agent who selects a policy implements its

[3]Put $\tilde{x}_i := x_i + \beta_P\theta$ and $\tilde{\theta} := (\beta_A - \beta_P)\theta$. Then $\tilde{x}^{(\tau)} = x^{(\tau)} + \beta_P\theta$, the principal's payoff is $a(\tilde{x}^{(\tau)} - \kappa)$ and the agent's is $a(\tilde{x}^{(\tau)} - \kappa + \tilde{\theta})$. The pivotal state is unchanged, a density is carried to a density by the linear change of variables, and (A1)–(A3) are imposed on $(\tilde{x}_1, \tilde{x}_2, \tilde{\theta})$, so that (A3) reads $\mathbb{E}[x_i + \beta_A\theta] < \kappa$.

recommendation[4] Direct mechanisms are menus indexed by types, $\tau \mapsto q_\tau$, with each type selecting its own policy. A direct mechanism satisfies incentive compatibility (IC) and individual rationality (IR) if

$$U_\tau(q_\tau) \geq U_\tau(q_{\tau'}) \qquad \text{for all } \tau, \tau' \in [0,1], \tag{IC}$$

$$U_\tau(q_\tau) \geq 0 \qquad \text{for all } \tau \in [0,1]. \tag{IR}$$

The revelation principle implies that, without loss of generality, we can focus on direct mechanisms that satisfy incentive compatibility and individual rationality.

The timing is as follows.

1. The principal posts a menu of acceptance policies.
2. The agent observes his type and selects a policy, or opts out.
3. The state realizes, independently of the type, and the selected policy is applied to it.
4. The recommended action is implemented and payoffs realize.

The principal's problem is therefore

$$\max_{(q_\tau)_{\tau\in[0,1]}} \int_0^1 V_\tau(q_\tau)\, d\tau \quad \text{subject to (IC) and (IR).}$$

3 The Optimal Menu

This section characterizes the solution to the principal's problem. We first solve the problem when the agent's type is known, which fixes the benchmark. We then show that the principal generally cannot achieve this first-best outcome once types are private information. The main result follows: the optimal menu consists of linear scoring rules governed by two parameters that satisfy a first-order condition, which we then interpret. The section closes with the solved menu for the admissions example. Proofs are deferred to a separate appendix.

[4]Two readings support this assumption: a centralized evaluation office that returns only the recommendation to the agent, or verifiable compliance, the policy prescribing a decision for every state and adopted policies being auditable. A school may not adopt another school's protocol and then reject the students it would admit. The assumption bites less than it seems. Write $U(\tau) \coloneqq U_\tau(q_\tau)$ for the payoff a type obtains from his own policy. A type who selected his own policy follows its recommendations voluntarily: after a recommendation to accept, his conditional payoff is $U(\tau)/\mathbb{E}[q_\tau] \geq 0$ under (IR), and after a recommendation to reject it is $\big(\mathbb{E}[x^{(\tau)} - \kappa + \theta] - U(\tau)\big)/\mathbb{E}[1 - q_\tau] < 0$ by (A3). What the assumption rules out is the double deviation of selecting another type's policy and inverting its recommendations. Section 5 shows that such deviations add exactly two inequalities, that both hold for every incentive compatible menu when $\max\{\mathbb{E}[(x_1 - x_2)^+], \mathbb{E}[(x_2 - x_1)^+]\} \leq \kappa - \mathbb{E}[\theta] - \max\{\mathbb{E}[x_1], \mathbb{E}[x_2]\}$, and that they hold with slack in the leading example.

Scoring rules and the private benefit of a policy. We begin with some preliminary analysis that rewrites the principal's and the agent's payoffs in terms of surpluses on match attributes and private benefits. For an acceptance policy q, let

$$S_i(q) \coloneqq \mathbb{E}\big[(x_i - \kappa)\, q(\omega)\big], \quad i = 1, 2, \qquad\qquad T(q) \coloneqq \mathbb{E}\big[\theta\, q(\omega)\big].$$

$S_i(q)$ is the surplus the policy generates on match attribute i, and $T(q)$ is the *expected private benefit* the policy delivers. Since $x^{(\tau)} - \kappa = \tau(x_1 - \kappa) + (1-\tau)(x_2 - \kappa)$, the payoffs of Section 2 can be rewritten as

$$V_\tau(q) = \tau S_1(q) + (1-\tau) S_2(q), \qquad U_\tau(q) = V_\tau(q) + T(q). \tag{1}$$

The agent's payoff from any policy is the principal's payoff plus the expected private benefit. The two parties therefore disagree about one thing only: how much private benefit a policy delivers.

A *linear scoring rule* with weight $\sigma \in \mathbb{R}$ and adjustment $\gamma \in \mathbb{R}$ is the policy

$$q^{\sigma,\gamma} \coloneqq \mathbb{1}\big\{\sigma x_1 + (1-\sigma) x_2 - \kappa + \gamma\theta > 0\big\}.$$

It scores the state by the weighted sum of the match attributes, with weight σ on the first attribute and $1-\sigma$ on the second, and takes the action when the score exceeds the threshold $\kappa - \gamma\theta$. The adjustment is the rate at which the threshold responds to the private benefit. A positive adjustment lowers the threshold when θ is high, in the agent's favor; a negative adjustment lowers it when θ is low, against him. In the admissions example, a negative adjustment makes the protocol more lenient toward applicants who need financial aid.

We highlight two instances of linear scoring rules. The *first-best policy* of type τ is $q^{\tau,0}$: it scores the state by the type's own match quality and ignores the private benefit. It is the policy the principal would choose for the type if she faced neither incentive nor participation constraints, since $q^{\tau,0}$ maximizes V_τ over all policies. The agent's own rule, the policy he would follow under full information and no regulation, is $q^{\tau,1}$. More generally, for every γ, the rule $q^{\tau,\gamma}$ is the unique maximizer over all policies of

$$V_\tau(q) + \gamma\, T(q) = (1-\gamma)\, V_\tau(q) + \gamma\, U_\tau(q), \tag{2}$$

because the integrand $(x^{(\tau)} - \kappa + \gamma\theta)\, q(\omega)$ is maximized state by state.[5] A rule with the agent's own weight, $\sigma = \tau$, and an adjustment in $[0, 1]$ is thus the policy that maximizes a weighted sum of the two parties' payoffs, with Pareto weight γ on the agent.

Finally, write $T^{\mathrm{fb}}(\tau) \coloneqq T(q^{\tau,0})$ for the expected private benefit of type τ's first-best policy. We call T^{fb} the *first-best benefits*: it records how much private benefit each type would obtain if the principal could impose the type's first-best policy.

[5] Uniqueness is up to π_0-null sets: the set $\{x^{(\tau)} - \kappa + \gamma\theta = 0\}$ has probability zero under (A2).

Known type benchmark. Suppose first that the principal knows the agent's type τ. Her problem is then a persuasion problem with a single receiver: choose a policy to maximize $V_\tau(q)$ subject to the agent's willingness to participate, $U_\tau(q) \geq 0$.[6] Call the type *free* if its first-best policy is acceptable to the agent, $U_\tau(q^{\tau,0}) \geq 0$, and *adverse* otherwise.

Proposition 1 (Known type)**.** *Suppose the principal knows the agent's type τ. Her optimal policy is the linear scoring rule $q^{\tau,\gamma}$ with the type's own weight and with adjustment $\gamma = \gamma^{\mathrm{kt}}(\tau)$, where*

$$\gamma^{\mathrm{kt}}(\tau) = \begin{cases} 0 & \text{if } \tau \text{ is free,} \\ \text{the unique } \gamma \in (0,1) \text{ with } U_\tau(q^{\tau,\gamma}) = 0 & \text{if } \tau \text{ is adverse.} \end{cases}$$

When the type is known, the principal scores the state by the agent's own criterion, $\sigma = \tau$. The only instrument she adds is a nonnegative adjustment, used to satisfy the participation constraint. A free type receives its first-best policy. An adverse type receives a rule that accommodates the private benefit just enough to make the agent indifferent to opting out. By (2), this rule maximizes a weighted sum of the two payoffs, and the adjustment is the weight on the agent that the participation constraint commands. In the admissions example at the parameters of Section 1.2, every school type is free: the board could impose each school's first-best protocol if it knew the school's priorities.

Is the first-best menu achievable? Now suppose the agent's type is private. A natural candidate menu offers every type its first-best policy. We call $(q^{\tau,0})_{\tau\in[0,1]}$ the *first-best menu*. It ignores the private benefit entirely and is the model's version of a blinding policy: every agent is scored by his own criterion, and no policy conditions on θ. If every type is free, the menu satisfies (IR), so the question is whether it satisfies (IC)

The answer turns on the first-best benefits. Consider a type τ contemplating the first-best policy of a nearby type r. Taking $q^{r,0}$ instead of $q^{\tau,0}$ changes his payoff by

$$\big[T^{\mathrm{fb}}(r) - T^{\mathrm{fb}}(\tau)\big] - \big[V_\tau(q^{\tau,0}) - V_\tau(q^{r,0})\big].$$

The second bracket is the loss in match quality, measured by his own criterion, from scoring with a neighbor's weights. It is nonnegative, since $q^{\tau,0}$ maximizes V_τ, and it is of second order in $|r-\tau|$ by the Envelope Theorem. The first bracket is the change in private benefit, and it is of first order wherever T^{fb} has a nonzero slope. So whenever the first-best benefits vary, some type profits from selecting a neighbor's scoring rule, and the first-best menu is not incentive compatible. Conversely, when T^{fb} is constant, every policy on the menu delivers the same private benefit, say $\bar{T}$, each type's payoff from any policy on it is $V_\tau(q) + \bar{T}$, and the agent ranks the policies exactly as the principal ranks them for his type: he selects his own first-best policy.

[6] Following the recommendation is then voluntary, by the argument in Section 2.

Proposition 2 (The first-best menu)**.** *Suppose every type is free. The first-best menu is incentive compatible if and only if T^{fb} is constant on $[0,1]$. In that case, it is the optimal menu.*

The second claim holds because the first-best menu gives the principal, type by type, the largest payoff any policy can give her; when it is feasible. The proposition says that blindness works in a knife-edge case and fails otherwise. In the admissions example, the first-best benefits are strictly decreasing in the school's weight on academics: academic-heavy cutoffs admit more applicants who need aid, because those applicants fall short on extracurriculars rather than on grades. Section 4 gives the closed form. Every school then gains by moving toward a neighbor with a more leadership-heavy cutoff, which is the distortion in Figure 1.

3.1 The optimal menu

Proposition 2 shows that the first-best payoff is generically not obtainable. What, then, does the principal's optimal menu look like? Our main result characterizes the solution:

Theorem 1 (The optimal menu)**.** *Under Assumption 1, the principal's problem has a unique solution $(q_\tau)_{\tau\in[0,1]}$, and it has the following form.*

(a) (Scoring rules.) *Every policy on the menu is a linear scoring rule: $q_\tau = q^{\sigma(\tau),\gamma(\tau)}$ with $\gamma(\tau) \in (-1,1)$, and the weight σ and the adjustment γ are continuous functions of the type.*

(b) (Pool and ramps.) *There is an interval $[\tau_-,\tau_+] \subseteq [0,1]$, possibly empty or a single point, such that:*

(i) every type in $[\tau_-,\tau_+]$ receives the same policy $\bar{q}$ and earns no rent, where $\bar{q}$ is the policy that maximizes the principal's payoff among the policies with balanced surpluses and zero rent,

$$\bar{q} \in \arg\max \big\{ V_\tau(q) :\ S_1(q) = S_2(q),\ U_\tau(q) = 0 \big\},$$

a policy that depends on the prior and the cutoff alone; every type outside $[\tau_-,\tau_+]$ earns a positive rent and receives a policy of its own;

(ii) on the lower ramp $[0,\tau_-]$, the weight and the adjustment satisfy

$$\frac{\sigma(\tau)-\tau}{1-\gamma(\tau)} = \int_0^\tau \frac{\gamma(s)}{1-\gamma(s)}\,ds; \tag{3}$$

(iii) on the upper ramp $[\tau_+,1]$, they satisfy

$$\frac{\tau-\sigma(\tau)}{1-\gamma(\tau)} = \int_\tau^1 \frac{\gamma(s)}{1-\gamma(s)}\,ds. \tag{4}$$

If the interval is empty, both (3) *and* (4) *hold on all of* $[0, 1]$.

Part (i) describes the *pool*: a set of types, possibly empty, who all receive one policy and are held to their opt-out payoff. With balanced surpluses, $S_1(q) = S_2(q)$, both payoffs from a policy are the same for every type, so the pool policy and the zero-rent condition do not depend on which types are pooled. The pool is where the participation constraint binds, and it can sit in the interior of the type space because, as noted in Section 2, (IR) is a constraint at every type. The two ramps are where types separate. Parts (ii) and (iii) are the first-order conditions of the problem, one for each ramp. Two consequences are immediate from the equations at the ends of the type space. If the pool does not reach the top, the integral in (4) is empty at $\tau = 1$, so $\sigma(1) = 1$: the highest type is scored by his own criterion. Likewise $\sigma(0) = 0$ if the pool does not reach the bottom, and both ends are undistorted when there is no pool. These conditions are analogous to the no distortion at the top condition in Mussa and Rosen (1978).

Corollary 1.1 (Blind menus)**.** *If T^{fb} is not constant, then $\gamma(\tau) \neq 0$ for some τ: the optimal menu conditions on θ. Consequently, every menu of policies that do not condition on θ is strictly suboptimal, including the first-best menu offered as is, with agents choosing freely among its policies.*

The corollary closes the argument that began with the first-best menu. If every adjustment were zero, the right-hand sides of the first-order conditions would vanish, every type on the ramps would be scored by his own criterion, and the menu would be the first-best menu, which is incentive compatible only when T^{fb} is constant by Proposition 2. Since the optimum is unique, no blind menu can achieve its payoffs. In particular, the outcome of offering the first-best menu and letting agents choose is itself an incentive compatible direct mechanism of policies that ignore θ, so it is strictly worse than the optimal menu.

Two distortions. Compare Theorem 1 with Proposition 1. With a known type there is no rotation, $\sigma(\tau) = \tau$, and the adjustment is nonnegative and used only for participation. With private types the theorem shows two new distortions. The weight is rotated, $\sigma(\tau) \neq \tau$: the principal scores the state as if the agent's criterion were a different one, which we call his *virtual criterion*. And adjustments can take either sign; in particular negative adjustments, taxes, never occur with a known type. Both are screening distortions. They arise from the type being private, not from the conflict over the private benefit by itself.

The virtual criterion. The first-order conditions can be read through the objective that each policy maximizes. Observe that, along the optimal menu, type τ's policy is the unique maximizer

over all policies of

$$V_{\sigma(\tau)}(q) + \gamma(\tau)\, T(q).$$

The principal evaluates match quality by the virtual criterion σ instead of the agent's true type. The scoring rule also has adjustment rate $\gamma(\tau)$ on the expected private benefit. With a known type, the virtual criterion is the agent's private type and the adjustment lies between zero and one (Proposition 1). On the optimal menu, the adjustment may be negative. Further, the virtual criterion is also distorted from the agent's type τ, and (4) quantifies this distortion: on the upper ramp, $\tau - \sigma(\tau) = (1 - \gamma(\tau)) \int_\tau^1 \gamma/(1-\gamma)$.

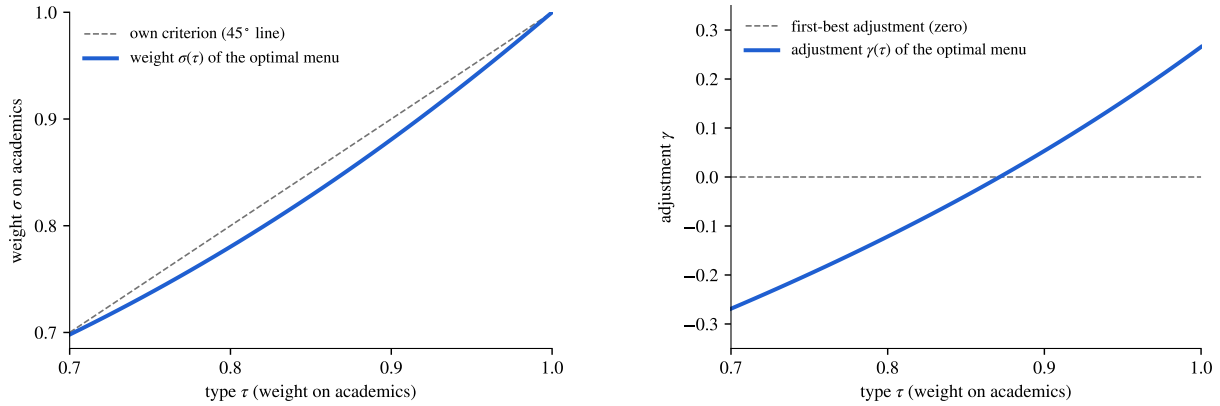
The distortion is the price of the rents that type τ's policy passes to the types above τ. Every policy on the menu is available to every type. By (1), a type $\tau' > \tau$ values a policy q more than type τ does exactly when $S_1(q) > S_2(q)$. Lowering the virtual criterion $\sigma(\tau)$ below τ reduces the surplus S_1 of type τ's policy relative to S_2. The types above τ then gain less from this policy, and the principal can leave them less rent. A unit of rent for type s costs the principal $\gamma(s)/(1-\gamma(s))$, the multiplier on the participation constraint in Proposition 1. This cost is positive where the adjustment is positive, and negative where the adjustment is a tax, since taxing type s less is a gain for the principal as well. The first-order condition (4) sums these costs over the types above τ. When the sum is positive, the principal lowers the virtual criterion below τ; when the sum is negative, she raises it above τ. At the top no type lies above, so $\sigma(1) = 1$. On the lower ramp, (3) sums over the types below τ. The two distortions are therefore linked: the virtual criterion of type τ is set by the adjustments of the types beyond τ.

3.2 The solved menu for the admissions example

Figure 2 shows the optimal menu for the admissions example, at the parameters of Section 1.2.[7] Panel (a) plots the weight $\sigma(\tau)$ of each school's policy against the school's own weight τ on academics, and panel (b) plots the adjustment $\gamma(\tau)$. The menu has no pool at these parameters: every school earns a positive rent, both balance equations hold on the whole type space, and both ends are undistorted, $\sigma(0.7) = 0.7$ and $\sigma(1) = 1$.

Every policy on the menu is a scoring rule with an adjustment strictly between -1 and 1, as Theorem 1 requires. Notice that the optimal menu features negative adjustments. Interpreting this feature: our main result says that, even in settings where the board of education does not prioritize

[7] The schools' weights on academics lie in $[0.7, 1]$ rather than $[0, 1]$. A change of variables maps this case into the model: with $x_1' \coloneqq x^{(1)}$ and $x_2' \coloneqq x^{(0.7)}$ as the match attributes and $s \coloneqq (\tau - 0.7)/0.3$ as the type, match quality is unchanged, $s\, x_1' + (1-s)\, x_2' = x^{(\tau)}$, and types are uniform on $[0, 1]$. The joint support of the new attributes is a parallelogram, which is why (A2) requires positivity only near the pivotal state. The figures keep the original scale.



(a) Weight

(b) Adjustment

Figure 2: The optimal menu for the admissions example. Panel (a): the weight $\sigma(\tau)$ of each type's policy against the type's own weight τ. Panel (b): the adjustment $\gamma(\tau)$; a negative adjustment holds applicants who need aid to a lower cutoff than other applicants, a positive adjustment to a higher one.

admitting low-income students, the optimal regulation it pursues may include admission protocols that are more lenient on those who need financial aid. The intuition is clear. In order to flatten the private benefits across types in the optimal menu, the board will use negative adjustments as an instrument to reduce the private benefits of leadership-heavy schools. Panel (a) shows that every interior school is assigned a virtual criterion below its own type, $\sigma(\tau) < \tau$: the board scores the school's applicants with more weight on extracurriculars than the school itself would. This is the sign that the first-order condition (4) predicts. Section 4 shows that this pattern, negative adjustments at the leadership-heavy end, positive adjustments at the academic end, and a virtual criterion that leans toward extracurriculars, holds whenever the first-best benefits are monotone.

4 Private Benefits and the Shape of the Menu

Theorem 1 characterizes the optimal menu by its first-order conditions. However, because priors over the three-dimensional state space are hard to characterize, we cannot easily sign the sign the distortions. Further, the sign of the rotation at a type τ_0 is determined by the adjustments of the types beyond him, and those adjustments are part of the solution. This section shows that one property of the prior allows us to obtain both signs: how the first-best benefits vary with the type. When the first-best benefits are monotone, every virtual criterion leans toward the same attribute when compared against the true type. The adjustments are then nonpositive at one end of the type space

and nonnegative at the other. The admissions example is the case in point. The section closes with a comparative static: how the principal's payoff changes when the match attributes become less informative about the private benefit. Proofs are deferred to a separate appendix.

4.1 Ordered private benefits

Recall that $T^{\text{fb}}(\tau) = T(q^{\tau,0})$ is the expected private benefit of type τ's first-best policy. Proposition 2 shows that the first-best menu is incentive compatible only when T^{fb} is constant. The natural next step is to study the problem when first-best benefits are monotone:

Definition 1 (Ordered private benefits)**.** Private benefits are *ordered* if the first-best benefits T^{fb} are monotone on $[0,1]$. We label the match attributes so that T^{fb} is nonincreasing, and we call x_2 the *favored attribute*.

Under this labeling, a type who places more weight on the favored attribute has a first-best policy that carries a weakly higher expected private benefit. Notice that ordered private benefits is a property of the prior and the cutoff; it can be computed without solving the optimal menu problem. In the admissions example the first-best benefits have a closed form. Let $m \coloneqq \mathbb{E}[x_2 \mid \text{aid}]$ be the mean extracurricular score among applicants who need aid, and let p be the fraction of applicants who need aid. For every type in the range of the example,

$$T^{\text{fb}}(\tau) = -c\,p\Big[(1-m) - \frac{\kappa - m}{\tau}\Big], \qquad \frac{d}{d\tau}T^{\text{fb}}(\tau) = -\frac{c\,p\,(\kappa - m)}{\tau^2} < 0. \tag{5}$$

Only the mean m enters the formula. The sign comes from $m < \kappa$: applicants who need aid fall short on extracurriculars, so a more academic cutoff admits more of them. The following result generalizes the observed patterns we see in the admissions example to settings with ordered private benefits:[8]

Theorem 2 (Ordered private benefits)**.** *Suppose Assumption 1 holds, private benefits are ordered with x_2 the favored attribute, and the optimal menu has no pool. Then:*

(a) (One-sided rotation.) *$\sigma(\tau) \le \tau$ for every τ: the virtual criterion of every type leans toward the favored attribute. An interior type with $\sigma(\tau) = \tau$ receives his first-best policy.*

(b) (Ordered benefits along the menu.) *$T(q_\tau)$ is nonincreasing in τ.*

(c) (Ends.) *$\gamma(0) \le 0 \le \gamma(1)$. The ends are undistorted, $\sigma(0) = 0$ and $\sigma(1) = 1$, as in Theorem 1.*

[8] In the example, θ equals $-c$ for an applicant who needs aid and 0 otherwise, where c is the cost of aid. The formula uses that x_1 is uniform and independent of (x_2, θ). It also uses $\tau \ge \kappa$, so that the first-best cutoff on x_1 lies inside $[0,1]$ for every value of x_2. The parameters of Section 1.2 give $p = 0.4$, $m = 1/3$ and $\kappa = 0.65$.

(d) (Strict versions.) *If T^{fb} is strictly decreasing, then $T(q_\tau)$ is strictly decreasing, and no interval of types receives first-best policies.*

If instead x_1 is the favored attribute, (a)–(d) hold with every inequality reversed. If T^{fb} is constant, the optimal menu is the first-best menu.

The no-pool hypothesis can be relaxed to a pool that consists of the lowest type alone. This is the regime of the admissions example when the cost of aid is large. In these cases, the optimal menu still features one-sided rotation and ordered private benefits. However, as expected, the bottom type's weight may now be distorted.

Theorem 2 shows that, when first-best benefits are ordered, the optimal menu inherits the ordering of private benefits from the first-best menu. The intuition is clear. Consider the first-best menu when first-best benefits are strictly decreasing: each type gains by taking the first-best policy of a type just below him. That policy puts more weight on the favored attribute, so it carries a higher expected private benefit. The loss in match quality is of second order. The optimal menu corrects the failure with both instruments. The adjustment $\gamma(\cdot)$ is negative at the bottom, where the first-best benefits are highest, and positive at the top, where they are lowest. A negative adjustment lowers the expected private benefit a policy delivers, and a positive one raises it. Part (b) says that the correction stops short of completely equalizing the order. The virtual criterion $\sigma(\cdot)$ of every type still exhibits lean toward the favored attribute. This lowers the rent his policy passes to the types above him. The two distortions are linked. The proof of part (b) rests on an identity: along the menu, the expected private benefit changes from one type to the next only where the virtual criterion departs from the type's own, and it falls where the virtual criterion lies below the type. Part (b) is therefore a consequence of part (a).

The solved menu of Figure 2 illustrates the statement of the theorem. The adjustment is negative at the leadership-heavy schools and positive at the academic-heavy schools. The virtual criterion of every interior school leans toward extracurriculars. Both ends are undistorted. One feature of the solved menu goes beyond the theorem. The private benefits delivered along the menu are nearly equal across schools, far flatter than part (b) requires. The theorem fixes the sign of such variation but not its magnitude.

4.2 Less informative attributes

The principal's main friction is the result of correlation between match attributes and θ. Indeed, suppose the attributes were uninformative about θ and the private benefit had mean zero. Then no acceptance policy conditional on the match attributes alone could deliver any private benefit. The

first-best menu would then be incentive compatible. With this in mind, this subsection studies how the principal's payoff changes as the correlation between match attributes (x_1, x_2) and θ changes.

Fix the marginal distributions of x_1, x_2 and θ, and let x_1 be independent of (x_2, θ). Normalize $\mathbb{E}[\theta] = 0$. The normalization makes every first-best policy carry zero expected private benefit when θ is independent of the attributes. Independence is therefore the natural benchmark. Further, fix a joint distribution H of (x_2, θ) with the given marginals. For $\eta \in [0, 1]$, let P_η be the prior under which x_1 is independent of (x_2, θ) and (x_2, θ) is drawn from

$$H_\eta \coloneqq (1 - \eta)\,(\text{product of the marginals}) + \eta\, H.$$

We call η the *dependence* between the attributes and the private benefit. Raising η changes nothing but the dependence. The distribution of (x_1, x_2) is the same under every P_η, and so are every first-best policy and every type's first-best value. What changes is the private benefit the first-best policies carry, in proportion to η:

$$T^{\text{fb}}_\eta(\tau) = \eta\, T^{\text{fb}}_H(\tau) \qquad \text{for every } \tau. \tag{6}$$

More dependence scales up the first-best benefits. If private benefits are ordered under H, they are ordered under every P_η. The dependence also has an information reading. For $\eta_1 < \eta_2$, the attribute x_2 under P_{η_1} is a garbling of x_2 under P_{η_2} as a signal about θ. It is the same attribute with probability η_1/η_2, and an independent draw from its marginal otherwise.

Write $V^{\text{fb}}(\tau) \coloneqq V_\tau(q^{\tau,0})$ for the first-best value of type τ, which does not depend on η. Write $B \coloneqq \int_0^1 V^{\text{fb}}(\tau)\, d\tau$ for the principal's first-best payoff, and $V(\eta)$ for her optimal payoff under P_η. The difference

$$C(\eta) \coloneqq B - V(\eta) \;\geq 0$$

is her loss relative to the first best, and we call it the *cost of screening*.[9] The results of this subsection use none of the structure of Section 4.1. They require neither ordered private benefits, nor uniform types, nor the absence of a pool.

Lemma 3 (Independence)**.** *Under independence, $\eta = 0$, the first-best menu satisfies (IC) and (IR), and the principal attains her first-best payoff: $V(0) = B$ and $C(0) = 0$.*

Under independence every type is free: his payoff from his first-best policy is $V^{\text{fb}}(\tau) \geq 0$. Every first-best policy also carries zero expected private benefit, so the first-best menu is incentive compatible by Proposition 2. Independence is the benchmark at which screening costs nothing. The next result compares two levels of dependence.

[9] If $\mathbb{E}_H[\theta \mid x_2]$ is nondecreasing in x_2, every type is free at every η. The principal could then impose the first-best policy on any known type, and $C(\eta)$ is exactly the cost of not knowing the type. A monotone likelihood ratio of x_2 in θ suffices.

Theorem 3 (Less dependence, type by type)**.** *Let* $0 \leq \eta_1 < \eta_2 \leq 1$ *and* $\lambda \coloneqq \eta_1/\eta_2$*. For every menu* $(q_\tau^2)_{\tau\in[0,1]}$ *that satisfies (IC) and (IR) under* P_{η_2}*, there is a menu* $(\tilde{q}_\tau^1)_{\tau\in[0,1]}$ *that satisfies (IC) and (IR) under* P_{η_1} *and gives the principal, at every type,*

$$V^{\mathrm{fb}}(\tau) - V_\tau^{\eta_1}(\tilde{q}_\tau^1) \;=\; \lambda\left[V^{\mathrm{fb}}(\tau) - V_\tau^{\eta_2}(q_\tau^2)\right], \tag{7}$$

where V_τ^η *denotes the principal's payoff under* P_η*.*[10]

Theorem 3 shows that, when x_2 becomes less informative about θ, one can construct a feasible menu that features reduced screening costs *for each type*. The proof is a mixing argument. The prior P_{η_1} is a mixture of the independence prior and P_{η_2}, with weights $1-\lambda$ and λ. The first-best menu is feasible under independence by Lemma 3. Mix it with the old menu, state by state and with the likelihood weights of the two components. The mixed menu is feasible under P_{η_1}, because (IC) and (IR) are linear in the policy. Its payoff at every type is the same mixture of the two payoffs, which is (7). The constructed policies are generally not scoring rules, and the constructed menu is generally not optimal. The immediate corollary of Theorem 3 compares the principal's aggregate screening costs:

Corollary 3.1 (The cost of screening)**.** V *is nonincreasing and concave in* η*, and* C *is nondecreasing and convex, with* $C(\eta_1) \leq (\eta_1/\eta_2)\,C(\eta_2)$ *for every* $\eta_1 < \eta_2$*. If* $C(\eta_2) > 0$*, then* $V(\eta_1) > V(\eta_2)$*.*

The cost of screening is governed by how much the attributes reveal about θ. By (6), the dependence scales the first-best benefits. The corollary says that the cost of screening rises convexly with that scale.

One can also make the construction deterministic. Suppose x_1 has no atoms and the old menu consists of scoring rules. Then each type's new policy can be chosen to take values in $\{0,1\}$. It departs from the type's first-best policy only where the old policy did, on a set of probability λ times the old one. Call a decision correct when it agrees with the type's first-best policy. Every correct decision stays correct. A positive mass of incorrect decisions is corrected whenever the old policy erred with positive probability. This is the sense in which less informative attributes let the principal depart from her preferred decision less often.

Two comparisons are not available. First, the constructed menu is not the optimal menu under P_{η_1}, so the theorem does not order the two optimal menus type by type. The principal maximizes an integral over types and can trade payoffs across them. Her payoff at a given type can therefore be lower under the new optimum if she decides to re-optimize across the available type-wise payoffs. Second, the theorem does not order the virtual criteria of the two optimal menus. A mixture of two

[10]The proofs of Theorem 3 and Corollary 3.1 allow the agent to invert a policy's recommendations as well as to follow them, so the results hold a fortiori under (IC) and (IR).

scoring rules is not a scoring rule. A scoring rule that delivers the same payoffs to a type as the mixture can be rotated even when neither rule in the mixture is.

5 Extensions

Two assumptions of Section 2 were made for convenience: an agent who selects a policy follows its recommendations, and types are uniformly distributed. This section relaxes each in turn. Proofs are deferred to a separate appendix.

5.1 Inverting recommendations

An agent who selects a policy might take the opposite action from the one it recommends: accept when the policy recommends rejection, and reject when it recommends acceptance. We say that the agent *inverts* the policy. The two actions divide the state space between them, so the payoff of type τ from inverting a policy q is $\mathbb{E}[x^{(\tau)} - \kappa + \theta] - U_\tau(q)$, the payoff from choosing $a = 1$ in every state less the payoff from following the policy. Inverting his own policy never pays: the first term is negative by (A3), and $U(\tau) \geq 0$ by (IR). What (IC) leaves open is the double deviation of selecting another type's policy and inverting it. Allowing it adds the constraints

$$U(\tau) \;\geq\; \mathbb{E}[x^{(\tau)} - \kappa + \theta] - U_\tau(q_{\tau'}) \qquad \text{for all } \tau, \tau' \in [0,1]. \tag{8}$$

The next result reduces this family to two inequalities.

Proposition 4 (Inverting recommendations)**.** *Let the menu $(q_\tau)_{\tau\in[0,1]}$ satisfy (IC) and (IR). It satisfies* (8) *if and only if*

$$U(0) + U(1) - \big[S_1(q_1) - S_2(q_1)\big] \;\geq\; \mathbb{E}[x_2 + \theta] - \kappa, \tag{9}$$

$$U(0) + U(1) + \big[S_1(q_0) - S_2(q_0)\big] \;\geq\; \mathbb{E}[x_1 + \theta] - \kappa. \tag{10}$$

Both inequalities hold for every menu that satisfies (IC) and (IR) if

$$\max\big\{\mathbb{E}[(x_1 - x_2)^+],\; \mathbb{E}[(x_2 - x_1)^+]\big\} \;\leq\; \kappa - \mathbb{E}[\theta] - \max\{\mathbb{E}[x_1], \mathbb{E}[x_2]\}. \tag{11}$$

The two inequalities involve the ends of the type space only: the rents of the two end types and the surplus differences of the two end policies. Inequality (9) is the constraint that the lowest type does not invert the highest type's policy, and (10) is the mirror. Under (IC) every other pair of types has more slack than one of these two, which is why the family reduces to them. Condition (11)

involves the prior and the cutoff alone. It holds in the admissions example,[11] so the assumption that agents follow recommendations is without loss there, and the menu of Section 3.2 remains optimal when inversion is allowed. The comparisons of Section 4.2 were proved with inversion allowed and need no condition.

When (11) fails, (9) and (10) join the problem as two constraints on the rents and the surpluses of the end types. If one of them binds at the solution, the balance equations of Theorem 1 are unchanged in the interior of each ramp, and the conditions at the ends of the type space acquire a multiplier. We do not characterize this case.

5.2 General type densities

Let the type have a continuous density g, strictly positive on $[0, 1]$. The reduction to direct menus and the incentive analysis of Section 3 do not use uniformity. It enters Theorem 1 in two places.

The first is the form of the balance equations. The cost of a unit of rent at type s is now weighted by the mass $g(s)$ of that type, and so is the rotation at type τ. On the upper ramp, (4) becomes

$$\frac{\tau - \sigma(\tau)}{1 - \gamma(\tau)}\, g(\tau) = \int_{\tau}^{1} \frac{\gamma(s)}{1 - \gamma(s)}\, g(s)\, ds, \tag{12}$$

and (3) changes in the same way. The reading of Section 3 is unchanged. The virtual criterion of a type is set by the adjustments of the types beyond him, each weighted by its mass.

The second is bunching. Under a uniform density, two distinct types receive the same policy only if both are pooled. Under a general density the menu may bunch: an interval of types outside the pool shares one policy and earns positive rent. On a bunch the balance equation cannot hold type by type, since the policy does not move with the type. It holds in integrated form, as in the ironing condition of Mussa and Rosen (1978): the two sides of (12) agree on average over the bunch, and their difference, averaged from one end of the bunch, has a fixed sign. Under the regularity conditions of Theorem 1, the rest of its structure survives. The type space consists of a pool, where the participation constraint binds, of ramps, on which types separate, and of bunches, and two bunches are never adjacent. Theorem 2 uses the uniform density through Theorem 1, and we have not extended it. The comparisons of Section 4.2 hold for any type density.

[11] After the change of variables in the footnote to Figure 2, the attributes are $x_1' = x_1$ and $x_2' = 0.7x_1 + 0.3x_2$. Then $\mathbb{E}[(x_1' - x_2')^+]$ and $\mathbb{E}[(x_2' - x_1')^+]$ are at most 0.055, $\mathbb{E}[x_1'] = 0.5$, $\mathbb{E}[x_2'] = 0.51$, $\kappa = 0.65$ and $\mathbb{E}[\theta] = -0.4c$, so (11) holds with slack for every cost of aid $c \geq 0$.

6 Conclusion

Organizations delegate decisions to agents who know better which criteria to apply, and they restrict what those agents may learn. This paper asks what the restriction should look like when the agent's criterion is private. The difficulty is that the characteristic the organization wants excluded is correlated with the attributes the agent may use. Blinding the characteristic is optimal only in a knife-edge case. Otherwise, the optimal regulation is a menu of linear scoring rules. The menu conditions on the excluded characteristic: thresholds adjust to it, in the agent's favor or against him. Each agent is also scored by a virtual criterion, which generally differs from his actual priorities over the match attributes. A first-order condition ties the two distortions together: the rotation of an agent's rule away from a Pareto-efficient solution is the price of the rents it passes to the other types.

When every type can raise his private benefit by overweighting the same attribute, the shape of the optimal menu reflects this property of the prior. Every virtual criterion leans toward the favored attribute, and the private benefits over the optimal menu retain the monotonicity of the first-best benefits. The cost of screening is governed by how much the attributes reveal about the characteristic. As match attributes become less correlated with the excluded characteristic, the principal suffers from lower screening costs when types are unknown. Further, we demonstrate that *negative adjustments* over the private benefits are a common feature of the optimal solution. In the admissions setting, while the board is indifferent to financial need, its optimal menu involves holding applicants who need aid to a lower cutoff in the leadership-heavy schools.

Two extensions are worth pursuing. The first is a simpler regulation. A menu on which every rule delivers the same private benefit is incentive compatible by construction. In the admissions example, the optimal menu delivers nearly equal private benefits across schools, which suggests that such a regulation loses little. When it captures most of the optimal menu's gain over blinding is a natural next question. The second is the agent's decision. We studied a binary action and two match attributes. Whether the virtual-criterion form of the solution survives richer actions or more attributes is open.

References

AGAN, A. AND S. STARR (2018): "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *Quarterly Journal of Economics*, 133, 191–235.

AMADOR, M. AND K. BAGWELL (2013): "The Theory of Optimal Delegation with an Application to Tariff Caps," *Econometrica*, 81, 1541–1599.

ANTONOVICS, K. AND B. BACKES (2014): "The Effect of Banning Affirmative Action on College Admissions Policies and Student Quality," *Journal of Human Resources*, 49, 295–322.

BERGEMANN, D., T. HEUMANN, AND S. MORRIS (2026): "Screening with Persuasion," *Journal of Political Economy*, 134, 570–625.

CANDOGAN, O. AND P. STRACK (2023): "Optimal Disclosure of Information to Privately Informed Agents," *Theoretical Economics*, 18, 1225–1269.

CHAN, J. AND E. EYSTER (2003): "Does Banning Affirmative Action Lower College Student Quality?" *American Economic Review*, 93, 858–872.

ELLISON, G. AND P. A. PATHAK (2021): "The Efficiency of Race-Neutral Alternatives to Race-Based Affirmative Action: Evidence from Chicago's Exam Schools," *American Economic Review*, 111, 943–975.

FRYER, R. G., G. C. LOURY, AND T. YURET (2008): "An Economic Analysis of Color-Blind Affirmative Action," *Journal of Law, Economics, and Organization*, 24, 319–355.

GUO, Y. AND E. SHMAYA (2019): "The Interval Structure of Optimal Disclosure," *Econometrica*, 87, 653–675.

HANCART, N. (2025): "The Optimal Menu of Tests," Working paper.

JULLIEN, B. (2000): "Participation Constraints in Adverse Selection Models," *Journal of Economic Theory*, 93, 1–47.

KOLOTILIN, A. (2018): "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics*, 13, 607–635.

KOLOTILIN, A., T. MYLOVANOV, A. ZAPECHELNYUK, AND M. LI (2017): "Persuasion of a Privately Informed Receiver," *Econometrica*, 85, 1949–1964.

KOLOTILIN, A. AND A. ZAPECHELNYUK (2025): "Persuasion Meets Delegation," *Econometrica*, 93, 195–228.

LEWIS, T. R. AND D. E. M. SAPPINGTON (1989): "Countervailing Incentives in Agency Problems," *Journal of Economic Theory*, 49, 294–313.

LI, Y. AND B. XU (2024): "Learning and Communication Towards Unanimous Consent," ArXiv:2405.18521.

MAGGI, G. AND A. RODRÍGUEZ-CLARE (1995): "On Countervailing Incentives," *Journal of Economic Theory*, 66, 238–263.

MILGROM, P. AND I. SEGAL (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70, 583–601.

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

MYERSON, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58–73.

STRACK, P. AND K. H. YANG (2024): "Privacy-Preserving Signals," *Econometrica*, 92, 1907–1938.