Screening Knowledge with Verifiable Evidence

Sulagna Dasgupta, Zizhe Xia [*]

Abstract

A principal seeks to screen an agent based on his demonstrable knowledge of a subject matter, modeled as a binary state. The agent learns about the state through two kinds of opposing verifiable signals, each kind providing evidence in favor of one of the states. A good quality agent has an evidence structure which is more informative than a bad quality one. In a symmetric setting, we show that under the optimal test, regardless of whether the agent can predict the state correctly, he is failed if the amount of evidence he is able to show is below a threshold. Conditional on providing evidence above this threshold, the agent is passed based on a simple True-False test – i.e., if and only if he gives the correct answer. We see this result as rationalizing a common test structure where test-takers are given credit for giving the correct answer only if they show a minimal amount of data, arguments, or steps, in support of their answer. We prove the results by identifying a connection to the optimal transport problem and leveraging it to show the existence of an appropriate virtual value function.

[*] We are most grateful to Ben Brooks, and Doron Ravid for their guidance and support. This project has also benefited tremendously from conversations with Emir Kamenica, , Phil Reny, Daniel Rappaport, Alex Frankel, Andreas Kleiner, Sarah Auster, Francesc Dilme and seminar participants at the University of Chicago and the University of Bonn. All errors are our own.

1 Introduction

Individuals are evaluated on their knowledge or expertise in a myriad of settings. Students take exams, job candidates are interviewed on their domain knowledge, consultants help firms make decisions and are often rewarded based on the ex-post accuracy of their advice, and so on. In many such knowledge-based evaluation schemes, accuracy of one's answers is not enough to earn rewards. For that, one must justify one's answers. Many exams give True-False or multiple choice tests to students, but specify that their answers must be justified in order for them to earn points. Similarly, it is typically not enough for a consultant assisting a firm in making a decision, to simply recommend a decision, even if it turns out to be correct in hindsight. He must provide exhaustive data and analysis to back up any recommendations he makes.

In this paper, we model the above testing setting as a problem of mechanism design with evidence. Specifically, the model is as follows. There is a binary state, unobserved by a test-taker/agent (he). He learns about it *only* through verifiable evidence. His *evidence type* is a vector with two real components – each component indicating the amount of evidence in favor of one of the states. The principal/test-designer (she) wants to design a pass/fail test. She observes the true state ex-post. She values both accuracy – how close the agent's belief is to the true state – and the amount of supporting evidence he possesses. We capture these features by imposing assumptions of, what we call, *accuracy monotonicity* and *evidence monotonicity*, on the principal's reduced form payoff function over the evidence space. This is her payoff from passing the agent. That from failing him is normalized to zero for both the principal and the agent.

We focus on the principal's optimal tests. Specifically, she can commit to a "test" – without loss, a pair of passing probabilities for each state, contingent on the agent's reported evidence. The objective is to maximize her ex-ante payoff, subject to the agent's incentive constraint which requires that it should be optimal for the agent to reveal all his evidence. The twist in the incentive constraints in our setting, vis-Ã -vis standard mechanism design settings is that, ours are "one-sided": The agent can hide evidence, but cannot manufacture it. Therefore, any evidence type can deviate only to its South-West, i.e., misreport any amount of evidence which is component-wise weakly lower. The set of possible misreports of a given evidence type is shown in Fig. 1.[1]

[1] The nature of allowed deviations is similar to Dziuda (2011), though our problem has commitment, unlike hers, and the setting is entirely different.

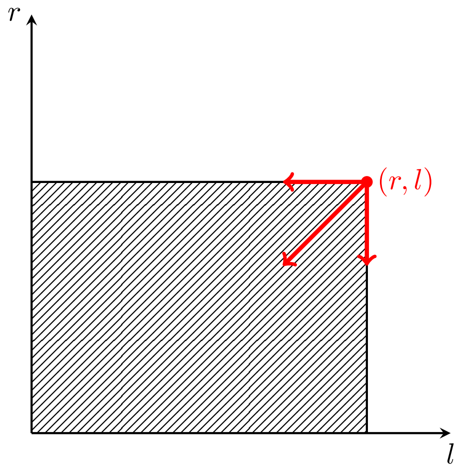


Figure 1: The shaded area depicts the set of misreports the evidence type (r, l) can make. Red arrows show a few examples of possible deviations.

The role of hard evidence in shaping incentive forces is as follows. Due to our assumption that each type of evidence favors one of the states, the agent's belief about any state is increasing in the favorable type of evidence and decreasing in the other one. This leads to positively sloped *isobelief curves* – curves along which the agent's belief remains the same. We show that, although each type has uncountably many directions of deviations available, it is sufficient to only consider deviations which are (1) horizontal (i.e., hiding only l), (2) vertical (i.e., hiding only r), or (3) "diagonal",[2] i.e., along an *isobelief curve* (hiding some of both l and r, while truthfully reporting one's belief). The relevant deviations are schematically represented by the red arrows in Fig. 1. The diagonal incentive constraints are our *evidence constraints*, in the language of the literature[3] – this is the set of constraints for which the restrictions on the directions of deviation, matter, in the following sense. Our analysis shows that if each type was allowed to deviate along its isobelief line in *both* directions, but restricted to only leftward and downwards deviations horizontally and vertically respectively, we would effectively be back in a standard mechanism design setting where information is "soft".[4] In other words, the unidirectionality of only the diagonal incentive constraints have a bite in our setting, not the horizontal and vertical ones.

Unidirectionality of incentive constraints has bit in our setting, precisely because of a novel trade-off in the knowledge screening setting we identify, which we call the *evidence/accuracy trade-off.* This is a trade-off the agent faces, due to the interaction of two forces in our model: the principal's preference for both accuracy and evidence, and verifiability of *all* of the agent's private information. On the one hand, the principal's taste for evidence pushes him towards showing all his evidence. But on the other, her taste for accuracy might make him want to exaggerate his knowledge (i.e.,

[2] We just call it *diagonal* for simplicity of terminology. Isobelief curves can have any positive slope and therefore a deviation along one of them can be in any South-West direction, not necessarily the 45° one.

[3] E.g., Vaidya (2023).

[4] That would be Dasgupta (2024)'s setting, who considers also considers the problem of screening knowledge, but allows for only belief-based – as opposed to evidence-based – screening.

the precision of his belief), which he cannot do without hiding some of his evidence. We call this the agent's *evidence/accuracy trade-off*. This trade-off leads to "unidirectionality" of our incentive constraints having bite in our setting, unlike many related settings of mechanism design with evidence (Celik, 2006; KrÃ€hmer and Strausz, 2024). Clearly, this is the force which also rules out *unraveling* – i.e., the agent's optimal strategy being revealing all his evidence, always.

We now describe our main characterization of optimal mechanisms. We first consider the case when the principal's relative preference for accuracy vis-a-vis evidence is strong enough so that there is no evidence type she ideally wants to accept in both states. In this case, in a symmetric setting, we show that the optimal test is simple True-False with an evidence threshold. This can be considered a natural generalization of the simple True-False test (which has been shown to be generically optimal for a large class of problems in Dasgupta (2024)), and is ubiquitous in the real world. In particular, the optimal test asks the agent to predict the state and passes him if and only if he is correct, as long as he provides a minimum level of supporting evidence. Examples of such tests are shown in Fig. 2 below. We see this result as providing a rationalization for the common test format where test-takers are rewarded based on the correctness of their answers, but only if they provide at least some amount of justification or reasoning for their answers, show a minimum number of steps – in case of mathematical or logical problems – or provide data in favor of their recommendations, in case of knowledge workers, and so on.

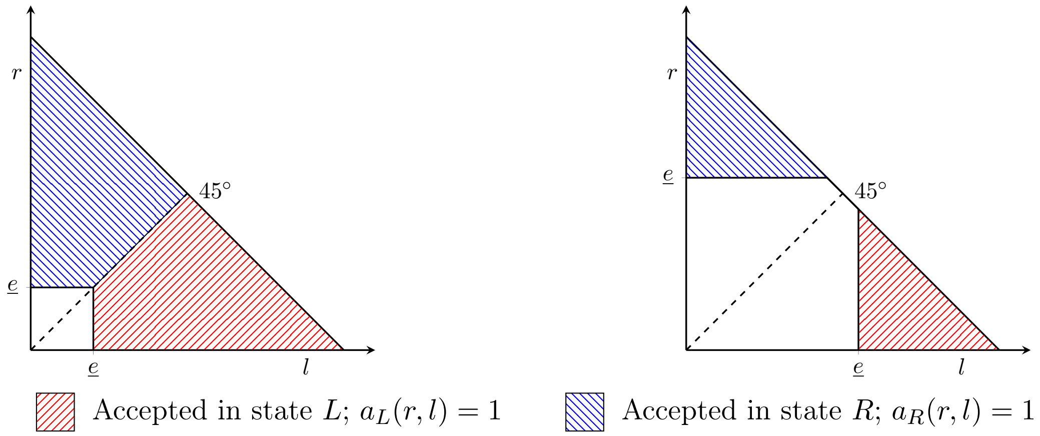

r
45°
$\underline{e}$
$\underline{e}$
l
r
$\underline{e}$
45°
$\underline{e}$
l
▨ Accepted in state L; $a_L(r,l) = 1$
▨ Accepted in state R; $a_R(r,l) = 1$

Figure 2: The state is $\omega \in \{R, L\}$. r (respectively, l) denotes the amount of evidence in favor of state R (respectively, L), shown by the agent. We assume the evidence space is $\{(r,l) \geq (0,0) : r + l \leq 1\}$. The figures show examples of mechanisms within our proposed optimal class, which are of the form $a_R(r,l) = 1(r \geq \max\{\underline{e}, l\}, a_L(r,l) = 1(l \geq \max\{\underline{e}, r\}$.

In an extension we relax the assumption of the principal not wanting to pass any evidence type in both states. In this case, under a regularity condition, we show that the optimal test takes a form which can be considered a further, natural generalization of the simple True-False test. In particular, the optimal test passes (respectively, fails) the test-taker regardless of his answer, if the amount of evidence provided is sufficiently high (respectively, sufficiently low), and passes if and only if his answer is correct, when the amount of evidence provided is intermediate.

Taken together, these results provide rationalization for commonly observed test structures, where students, job interview candidates etc. are rewarded for the correctness of their answers, only if they provide a minimum amount of reasoning/justifications for their answers, and conversely, sometimes they are also rewarded if they show enough reasoning, but are unable to arrive precisely at the correct answer.

1.1 Our Methodology

The standard mechanism design toolkit does not apply in our setting due to two issues – the multi-dimensionality of evidence types and the one-sidedness of the incentive constraints. Consequently, we use a duality approach, certifying the optimality of the proposed optimal mechanism via directly constructing the multipliers on the incentive constraints. We do this by identifying a novel connection of our problem to the optimal transport problem.

We observe that the multipliers on the incentive constraints serve as a transport plan that shifts the principal's virtual values along the binding constraints. Our proposed optimal mechanism is deterministic. So, to certify the optimality of such a mechanism, we have to construct multipliers so that the virtual values are positive for evidence types that are always passed by the mechanism and negative for those always failed. This involves "transporting" the virtual values through multipliers to create the correct signs. The existence problem of such multipliers is then converted into the existence problem of certain transport plan. This transport plan must satisfy a directional constraint that says virtual values can only be shifted in the North-East directions, as a result of the one-sided incentive constraints. We then apply a classic result from the statistics literature – Strassen (1965)'s theorem – to show that such a transport plan exists. In this regard, our methods share a connection with those of Haghpanah and Hartline (2021), though their results do not apply to our setting.

1.2 Related Literature

We provide an overview of the related literature here, relegating a more detailed discussion to Appendix A.

This paper mainly contributes to three distinct strands of literature – screening knowledge, mechanism design with evidence, and multi-dimensional mechanism design. Less directly, it also contributes to the literature on mechanism design with verificaiton and that on test design. Within the literature on screening knowledge, the closest to our paper are (Dasgupta, 2024) and Deb, Pai and Said (2023). Within that on mechanism design with evidence, our work relates most closely to Sher and Vohra (2015), Celik (2006), KrÃ€hmer and Strausz (2024), Dasgupta, Krasikov and Lamba (2022) and Vaidya (2023). Finally, within the multi-dimensional mechanism design literature, our work is closest to Haghpanah and Hartline (2021), which in turn builds on Carroll (2017) and Cai, Devanur and Weinberg (2019).

2 Model

The model features a principal (she) who is a test-designer and an agent (he) who is a test taker. The principal tests the agent on his knowledge of some binary state to decide whether to pass or fail him. There are no monetary transfers.

Our main model is in a reduced form. We provide a microfoundation for our model and its features in Section 5.

Evidence and Learning There is a (binary) unknown state $\omega \in \{R, L\}$ with a common prior that R and L are equally likely. The principal does not know about the state when she designs the test. She learns the state ex post, and can condition the pass-fail decision on the state.

All learning is verifiable. The agent learns about the state only through verifiable evidence. There are two kinds of evidence, each favoring one of the states – in a sense to be made precisely shortly. Let $r, l \geq 0$ denote the respective *amount* of evidence favoring each state. The pair (r, l) is the agent's *evidence type* – his only private information.

Interpretation. The structure fits several applications. For example, a consultant or election forecaster trying to learn about a *state* – the right decision for the client, or the election outcome – may be able to collect data both for and against each possibility under consideration. A student may be able to come up with several arguments both for and against a given statement in an exam, without knowing if it is True or False. He may also be able to derive a few steps of a mathematical problem without knowing the correct answer. These are all examples of an agent learning the state partially through contradictory sets of evidence.

Resource constraint. The set of all possible evidence types is

$$E := \{(r, l) \geq 0 : \phi(r, l) \leq 1\}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$ is a function which is symmetric, convex, and strictly increasing in each argument.[5] $\phi(r, l) \leq 1$ is the aggregate resource constraint of evidence generation. For example, a student has time constraints in an exam which puts a cap on the number of steps he can derive or the arguments he can give in favor of either of the possible answers; a consultant may have time and other resource constraints on the amount of data he can collect in favor of either of the options he recommends.

[5] By *symmetric* we mean, $\phi(r, l) = \phi(l, r), \forall (r, l)$.

Evidence distribution. The evidence types are distributed over E according to some CDF $F(r, l \mid \omega)$ in state ω with continuous and positive density $f(r, l \mid \omega)$. We assume this distribution is symmetric, that is, $f(r, l \mid R) = f(l, r \mid L), \forall (r, l) \in E$. Let $f(r, l) := f(r, l \mid R) = f(l, r \mid L)$ going forward.

Contradictory nature of evidence. We capture the contradictory nature of the two kinds of evidence, in the following way. Knowing the evidence distribution, the agent forms posterior beliefs using Bayes rule. Let $p(r, l) := \Pr(R \mid r, l) = \frac{f(r,l)}{f(r,l)+f(l,r)}$ denote his posterior belief that the state is R. We want to capture the fact that r favors state R and l favors state L, by assuming that p increases in r and decreases in l. This is equivalent to the following assumption on the distribution of evidence.

Assumption 1. *The likelihood ratio $\gamma(r, l) := \frac{f(r,l)}{f(l,r)}$ strictly increases in r and strictly decreases in l.*

Preferences We normalize the principal's utility of failing the agent to zero, and let $u(r, l \mid \omega)$ be that from passing an agent of type (r, l), when the state is ω. Assume that $u(r, l \mid \omega)$ is again symmetric, that is, $u(r, l \mid R) = u(l, r \mid L), \forall (r, l) \in E$. Going forward we use the notation $u(r, l) := u(r, l \mid R) = u(l, r \mid L)$.

Principal values evidence *and* accuracy. We assume our principal exhibits a preference for both accuracy and evidence. We state this in terms of her *derived* payoff function $\tilde{u}$, over the agent's belief (capturing accuracy) and supporting evidence.

Let $\tilde{u}(p, e|\omega)$ denote the principal's payoff in any state ω, as a function of the agent's belief p that the state is ω, and *supporting* evidence he shows in favor of ω, i.e. if $\omega = R$, $e = r$ and if $\omega = L$, $e = l$. Clearly,

$$\tilde{u}(p_0, e|R) = u(e, p^{-1}(p_0; r)), \; \forall \, p_0 \in \left[\min_{(r,l)\in E} p(r, l), \max_{(r,l)\in E} p(r, l) \right], r \in [0, 1]$$

By assumption 1, the above is well defined. We can see that by symmetry of u, $\tilde{u}$ does not depend on ω. So we drop it from its arguments, going forward.

We capture the feature that the principal prefers more accurate beliefs and more evidence, by assuming that in each state, for a fixed level of supporting evidence shown, she prefers a higher belief, and similarly, for a fixed belief, she prefers higher amounts of supporting evidence being shown. Mathematically,

Assumption 2. 1. *(Accuracy Monotonicity) For any $p' > p, e$, $\tilde{u}(p', e) \geq \tilde{u}(p, e)$.*

2. *(Evidence Monotonicity) For any $p, e' > e$, $\tilde{u}(p, e') \geq \tilde{u}(p, e)$.*

It is easy to show, that Assumptions 2.1-2 on $\tilde{u}$, taken together, are equivalent to Assumptions 3.1-2 below, on u, taken together. We show this formally in Appx. B.1.

Assumption 3. *1. (Accuracy Monotonicity′) For any $l' < l$ and $r' > r$, $u(r', l) > u(r, l)$ and $u(r, l') > u(r, l)$.*

2. (Evidence Monotonicity′) For any $(r', l') \geq (r, l)$ with $p(r', l') = p(r, l)$, $u(r', l') > u(r, l)$.

For simplicity, in the main body of the paper we also assume that the principal's preferences are such that there is *no* type she wants to accept in *both* states. We partially relax this in the Extension section 6, and show how the class of optimal mechanisms changes in that case. Formally:

Assumption 4. *For all $r, l \in E$, $u(r, l) \leq 0$, or $u(l, r) \leq 0$, or both.*

Agent's preferences. The agent always wants to pass. Consequently, we normalize his payoff from passing to 1 and that from failing to 0. Both the principal and the agent are expected utility maximizers.

In Section 5, we provide a micro-foundation for the above reduced form model, where the agent is of either good or bad quality, and the principal wants to pass only the good quality agent.

Mechanisms Next, we describe our universe of mechanisms, and our notion of incentive compatibility, which is "one-sided", due to verifiability of evidence.

By a revelation principle by Bull and Watson (2007*b*), it suffices to consider direct mechanisms that incentivize full disclosure of the agent's evidence.[6] Hence we can, without loss, define our universe of mechanisms as follows.

Definition 1 (Mechanism)**.** A *mechanism* is a pair of functions $(a_R, a_L) : E \to [0, 1]^2$ that maps the agent's reported evidence to the probabilities of passing when the state is R and L respectively.

The principal chooses a mechanism that is incentive compatible in the following sense, to maximize her expected payoff.

Definition 2 (Incentive compatibility)**.** A mechanism is *incentive compatible* (IC) if it is optimal for the agent to fully disclose his evidence, that is,

$$\begin{aligned} a_R(r,l)p(r,l) + a_L(r,l)(1-p(r,l)) \geq a_R\left(r',l'\right)p(r,l) + a_L\left(r',l'\right)(1-p(r,l)), \\ \forall (r,l), \left(r',l'\right) \in E, \left(r',l'\right) \leq (r,l). \end{aligned} \tag{1}$$

The above IC constraints are "one-sided" – the agent can only misreport an evidence type that is

[6] To be precise, the revelation principle in Bull and Watson (2007*b*) says that it is without loss to consider direct mechanisms that incentivize truthful reporting of the agent's belief and full disclosure of evidence. Since the agent learns about the state only through evidence, it suffices to incentivize full disclosure of evidence.

component-wise weakly lower than the evidence he possesses. In other words, he can hide evidence he already has, but he cannot manufacture it.

Timing Nature draws ω. The principal chooses a mechanism (a_R, a_L) without observing ω. The agent's evidence type (r, l) given ω is then realized according to $F(r, l \mid \omega)$ and privately observed by the agent. He then decides what evidence $(r', l') \leq (r, l)$ to reveal.[7] The state ω is then publicly revealed. The mechanism then passes or fails the agent based on the reported evidence and the realized state.

3 An Example

In this section we present an example to illustrate the role of incentive issues and hard evidence in this screening problem.

Setting A firm (she) – the principal – wants to design a process (a mechanism) to decide on retaining or not a consultant (he) – an agent – based on his advice. He advises her on whether to invest in a certain project. Hence, our unknown binary state is whether, in hindsight, investing was the right decision (say, state R) or not (say, state L). Ex-ante, both decisions are equally likely to be correct, i.e., the prior over the state is one half.

The consultant's *quality* can be $G(ood)$ or $B(ad)$ with equal probability. We assume the firm's payoff from retaining a consultant of $G(good)$ quality is 1, and that from retaining one of $B(ad)$ quality is $-u_B$, where we assume $u_B \geq 1$. Her payoff from not retaining him is normalized to zero. Quality and the state are independent. We assume quality is unobserved by both the firm and the consultant.

Agent's learning The consultant learns about the state by collecting data – stakeholder feedback, financial and other internal reports of the firm etc. Some of this data offers support to investing being the right decision (evidence type r, in the language of our model), and some supports not investing (evidence type l). We assume the state is publicly revealed ex-post – ex post, it becomes clear to all parties if investing was the right decision.

We would model his data gathering process as a sequence of m i.i.d. experiments on the state, each producing one of three outcomes – a unit of r, a unit of l, or no evidence at all $(\emptyset)$. Going forward, with a slight abuse of notation, we use r and l to refer to the number of each type of evidence collected. We refer to the tuple (r, l) as *evidence*. With this experiment, the set of possible evidences – the *evidence space* – is:

[7] The agent's participation constraint is automatically satisfied, since not participating in the mechanism (i.e., not taking the test) guarantees a payoff of zero.

$$E := \{(r,l) \in (0 \cup \mathbb{N})^2 : r + l \leq m\}$$

Let β_q be the probability of the consultant of quality $q \in \{G, B\}$ generating any evidence from any such experiment, and let η_q be that of getting the "right" kind of evidence, conditional on generating any. The experiment is described by the table below, where rows represent outcomes, columns represent states and cell entries represent the conditional probability of each outcome, given each state.

	R	L
r	$\beta_q \eta_q$	$\beta_q (1 - \eta_q)$
l	$\beta_q (1 - \eta_q)$	$\beta_q \eta_q$
$\emptyset$	$1 - \beta_q$	$1 - \beta_q$

Table 1: Evidence distribution of each i.i.d. experiment. Rows represent outcomes, columns represent states and cell entries represent the conditional probability of each outcome, given each state.

With the above experiment, the evidence collected by each type follows the multinomial distribution, the probability mass function of which is given by:

$$f_q(r,l|R) = \frac{m!}{r!l!(m-r-l)!} \times (\eta_e)^r (1-\eta_e)^l \beta_e^{l+r} (1-\beta_e)^{m-(r+l)} \tag{Multinomial}$$

By symmetry, $f_q(r,l|L) = f_q(l,r|R)$, and we use $f_q(r,l)$ to denote $f_q(r,l|R)$ going forward.

We assume the $G(ood)$ type is both more *likely* to generate evidence and more likely to generate the *correct* kind of evidence, than the $B(ad)$ type. This is formalized as follows.

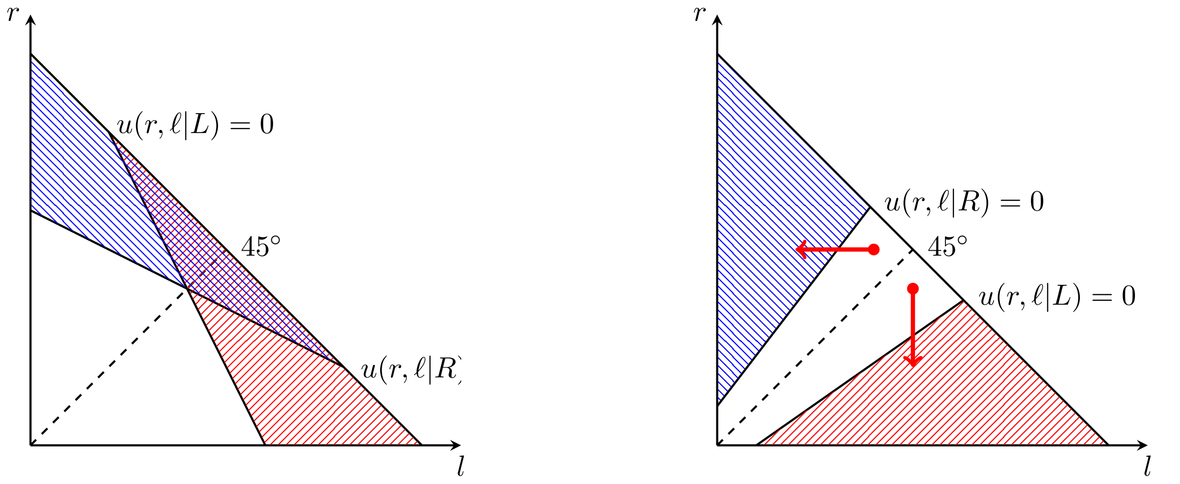
Assumption 5. *We assume $\eta_G \geq \eta_B$ and $\beta_G \geq \beta_B$.*

Interim quality The firm wants to design a mechanism $a_R, a_L : E \to [0,1]^2$ to maximize the ex ante expected quality of the retained consultant. His ex ante quality is the expectation of his *interim* quality – his expected quality given his evidence and the state. Algebra shows that is given by, using our notation from the model section:

$$u(r,l) = u(r,l|R) = \frac{f_G(r,l) - u_B f_B(r,l)}{f(r,l)} \tag{$u(r,l)$}$$

Note that the consultant does not observe his own quality, so the distribution of evidence, unconditional on quality, is given by the simple average of f_G and f_B, which we denote by f, i.e., $f := \frac{1}{2} f_G + \frac{1}{2} f_B$. Let its CDF be denoted by F.

First Best We first describe the firm's optimal test if she could observe the consultant's evidence but not his quality – which we call her *first best* – and describe its incentive violations.



(a) First best is implementable

(b) First best violates incentive constraints

Figure 3: The blue and red regions depict set of accepted types under the first best, in states R and L respectively. In case (a), the first best is implementable – i.e., incentive constraints do not bind – but not in case (b). In case (b), the red arrows show possible deviations, if the first best mechanism is offered.

The first best test is clearly given by $a^{fb}_{\omega}(r,l) := 1(u(r,l|\omega) \geq 0), \omega \in \{R, L\}$, where $a^{fb}_{\omega}(r,l)$ denotes the passing probability of evidence (r,l) under this mechanism, when the true state is ω.

By equation ($u(r,l)$), the $u(r,l)=0$ curve – above which all evidence types are retained, in state R – is given by the following straight line:

$$\frac{f_G(r,l)}{f_B(r,l)} = u_B, \text{after taking log on both sides and some algebra,}$$

$$r\left(\ln\left(\frac{\frac{\beta_G}{1-\beta_G}}{\frac{\beta_B}{1-\beta_B}}\right)+\ln\left(\frac{\eta_G}{\eta_B}\right)\right)+l\left(\ln\left(\frac{\frac{\beta_G}{1-\beta_G}}{\frac{\beta_B}{1-\beta_B}}\right)-\ln\left(\frac{1-\eta_B}{1-\eta_G}\right)\right)-m\ln\left(\frac{1-\beta_B}{1-\beta_G}\right)=\ln u_B \tag{FB-line}$$

Clearly, exchanging r and l in the above equation gives the curve *to the right of* which all evidence types are retained in state L. It is a mirror image of the above straight line, reflected across the 45° line.

Examples of first-best acceptance regions of the evidence space are shown in Fig. 3. They are shaded in red and blue in states R and L respectively.

Unraveling Now we would see that "unraveling" happens in this setting – i.e., the consultant of *every* evidence type has the incentive to show all his evidence – if and only if the Good and Bad quality consultants differ more in terms their evidence generation capability, than on the accuracy of that evidence.

Clearly, under our assumption 5, (FB-line) is strictly positively sloped if and only if:

$$\frac{\frac{\beta_G}{1-\beta_G}}{\frac{\beta_B}{1-\beta_B}} < \frac{1-\eta_B}{1-\eta_G} \tag{nontriviality}$$

When this condition is not met, they are negatively sloped, and we get the case depicted in Fig. 3a. Note that under the mechanism $\{a_\omega^{fb}\}_{\omega\in\{R,L\}}$, no evidence type (r,l) strictly prefers an allocation in its south west quadrant, i.e., in $\{(r',l') : r' \leq r, l' \leq l\}$. Hence, if the condition (nontriviality) is violated, the first-best is implementable.

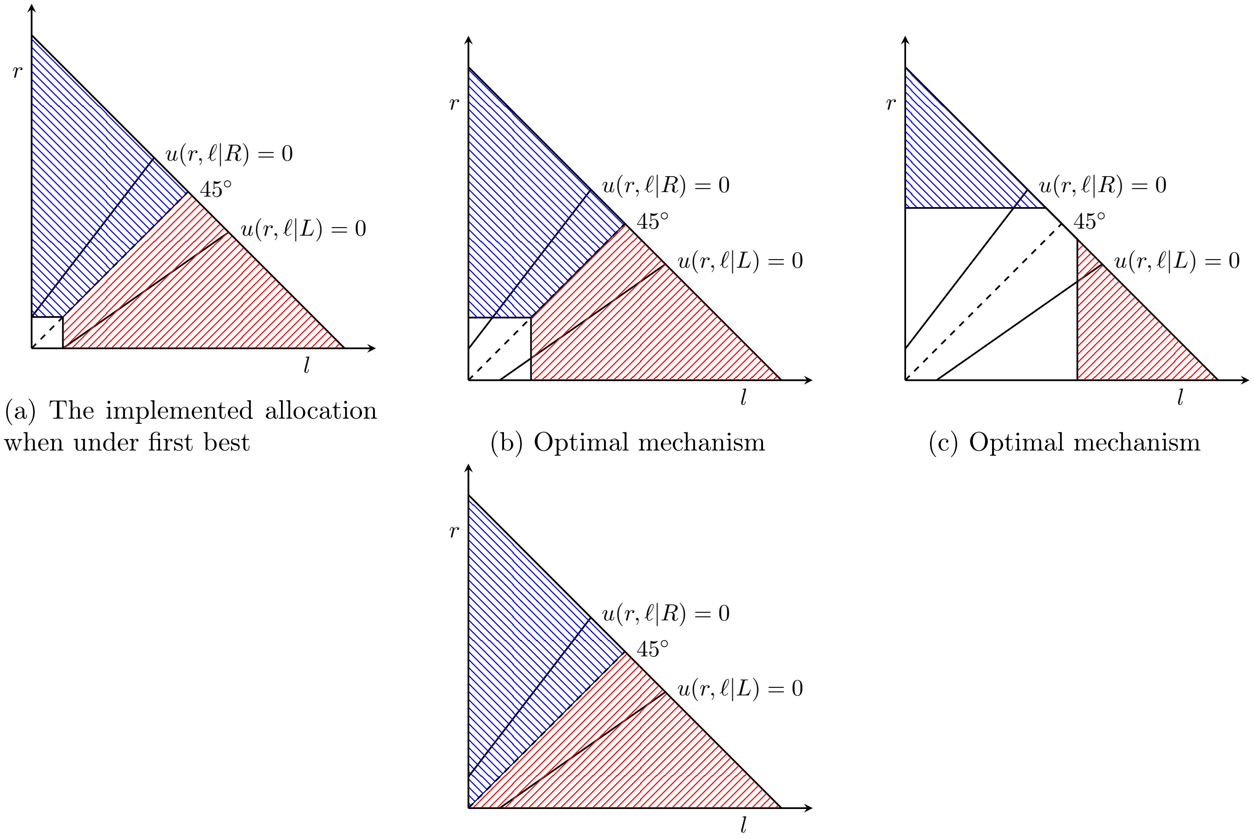
Intuitively, (nontriviality) means that, in a sense, the “difference” of accuracy between the Good and Bad types (captured by η_G and η_B) is *more* than that of their capacity to generate evidence (captured by β_G and β_B). Violation of this condition would, therefore, lead to the *quantity* of evidence acting as a stronger signal of better quality, than its *accuracy*. The only reason the agent might have to conceal any of his evidence, in our model, is a desire to come across as having a more accurate belief than he really does. If quantity of evidence is a better indicatory of quality to the principal, than its accuracy, this force is not present. In that case, the agent has no imperative to hide any of his evidence. Therefore unraveling occurs, which means, the first-best is implementable.

Role of incentives When (nontriviality) is satisfied, the $u(r,l|\omega) = 0$ lines are positively sloped. The first best, in this case, is depicted in Fig. 4b. Note that if this mechanism is offered, the deviations depicted by arrows occur. This implements the mechanism shown in Fig. 4a. The evidence types in the little square at the bottom are never retained, but due to the verifiable nature of evidence, they can’t deviate to a strictly better allocation, as all such allocations require strictly more of either l or r, which they can’t provide.

True-False with an evidence threshold. By the above reasoning, offering the first-best mechanism, is equivalent to offering the mechanism shown in Fig. 4a. But note that generically, the firm can do better than this mechanism – by choosing $\underline{e}$ optimally, given this mechanism, i.e.:

$$\underline{e} \in \arg\max_{e} \int_{r\geq\max\{e,l\}} u(r,l)F(dr,dl)$$

Examples of this are shown in Figures 4b and 4c. In the main body of the paper we show that this is, in fact, the best the principal can do, in a general class of environments, even though she can choose from a rich universe of mechanisms, including ones employing randomization. Under this mechanism, the agent is accepted if and only if he predicts the state correctly, conditional on



(a) The implemented allocation when under first best

(b) Optimal mechanism

(c) Optimal mechanism

(d) Optimal mechanism without hard evidence

Figure 4: The class of optimal mechanisms with and without verifiability

producing *supporting* evidence (i.e. r if he predicts R, etc.) of at least $\underline{e}$. Due to this structure, we call our optimal class of mechanisms, *True-False with an evidence threshold.*

Role of hard evidence. Note that, if we did not have verifiability, offering the first best mechanism would not have resulted in the implementation of the mechanism shown in Fig. 4a but the one in Fig. 4d – the familiar simple True-False – rationalized as optimal in a large class of settings by Dasgupta (2024), in the absence of hard evidence. Clearly, the our proposed optimal mechanisms of Figures 4b and 4c would not be implementable without verifiability, and offering either of those would lead to the mechanism in Fig. 4d being implemented.

4 Main results

In this section, we provide our main characterization of optimal mechanisms as ones that reward test-takers if and only if they give the correct answer, conditional on providing a minimum level of supporting evidence. Before that, we highlight the novel evidence/accuracy trade-off that arises in our setting, and the role of verifiability.

4.1 Unraveling and the evidence/accuracy trade-off

In this subsection we explain the central *evidence/accuracy trade-off* faced by the agent in our problem and use it to show why *unraveling* – revealing of all evidence – does not occur in our model. We also provide alternative natural conditions for unraveling to occur in this setting.

A natural question to ask may be, that since more evidence is preferred by the principal, why doesn't the agent simply reveal all his evidence, i.e. why doesn't *unraveling* always happen? The answer lies in the interaction of two forces in our model: (1) the principal's preference for both accuracy and evidence, and (2) the absence of any "soft" – i.e., unverifiable – information. In particular, the agent wants to come across as both precisely informed *and* possessing as much evidence as he can possibly show. But because he possesses *only* verifiable information, *if* he wants to pretend to have a more precise belief than he actually has, there is no way for him to do that other than showing less of his evidence than he actually has. We call this the agent's *evidence/accuracy trade-off.*

The aforementioned trade-off leads to the fundamental difference between our question, and related mechanism-design-with-evidence questions asked in the literature so far[8]: the directions of deviations in our model are both type-dependent and non-obvious. Modeling of partial verifiability of private types as a restriction on the directions of deviations is not new. However, in many standard mechanism design problems such as selling (Celik, 2006; KrÃ€hmer and Strausz, 2024), or procurement (KrÃ€hmer and Strausz, 2024), the direction in which the agent would *want* to deviate, regardless of his type, is obvious – a buyer always wants to understate his willingness to pay, a producer applying for a tender wants to exaggerate his costs, etc. Consequently, a restriction on

[8] To the best of our knowledge.

the direction of deviations often fails to have a bite, at least in reasonably "regular" environments – leading to the same optimal mechanisms as with no such restrictions (Celik, 2006; KrÃ€hmer and Strausz, 2024). In contrast, in our model, it is not clear how the agent resolves the evidence/accuracy trade-off, *even* given his evidence type, because which type of deviation would benefit him, if any, depends on the principal's relative preference for evidence vs accuracy.

A particular instance of that relative preference is when the principal has an "extreme" preference for evidence over accuracy, which – unsurprisingly – leads to unraveling. We formalize this below.

Definition 3 (Unraveling)**.** We say *unraveling* occurs, when there is an optimal solution to the principal's problem in which no incentive constraint binds.

Proposition 1 (Unraveling)**.** *Unraveling occurs under either of the following alternative settings.*

- *Modify Assumption 2 so that $\tilde{u}$ remains strictly increasing in e but does not depend on p. This is equivalent to u remaining strictly increasing in r, but not depending on l. In this case unraveling occurs.*
- *If instead, u is increasing in both r and l, then also unraveling occurs.*

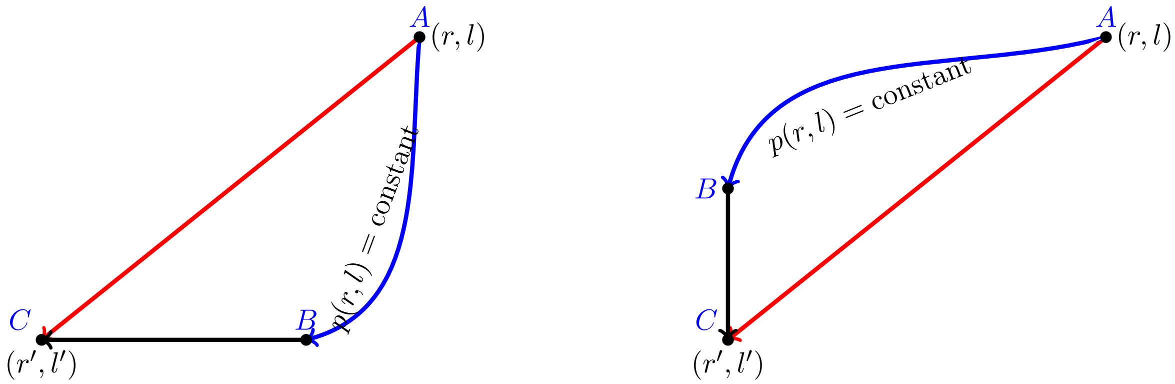
The first bullet point above is obvious – if the principal cares only about evidence, our problem is trivial: first best is always implementable. Combined with the second point, the above observation gives a sense of how strong we need the principal's relative preference for accuracy vis-a-vis evidence to be, for the evidence/accuracy trade-off to have any bite and therefore the problem to be non-trivial. In particular, we need her payoff to be either strictly decreasing or non-monotonic, in the wrong kind of evidence (recall that $u(\cdot,\cdot)$ is her payoff in state R, so l is the wrong kind of evidence). If this relative preference is mild enough so that her payoff weakly increases, not only in the correct but also the wrong kind of evidence, we must have unraveling.

The proof is simple and follows from the $u(r,l) = 0$ curve becoming weakly negatively sloped, under the assumptions of Observation 1. It is omitted for brevity.

4.2 Simplifying incentive constraints

In this subsection we highlight the fact that, although each evidence type has uncountably many directions of deviations available, it is sufficient to only consider deviations which are (1) horizontal (i.e., hiding only l), (2) vertical (i.e., hiding only r), or (3) "diagonal", i.e., along an *isobelief curve* (hiding some of both l and r, while truthfully reporting one's belief). This simplification of incentive constraints is instructive, as it provides an intuitive decomposition of such constraints on deviating across *beliefs*, which are "soft", or *bidirectional*, and those on deviating across *amounts of evidence* – while *truthfully* reporting the belief – which are "hard", or *unidirectional.*

In particular, we show that an IC $(r,l) \to (r',l')$ binds, as shown in Figures 8 below, if and only if corresponding diagonal and horizontal ($(r,l) \to (r',l'')$ and $(r',l'') \to (r',l')$), or diagonal and vertical ($(r,l) \to (r',l'')$ and $(r',l'') \to (r',l')$), as the case may be, pairs of IC's also bind (See Figure. 8). The details are relegated to Appendix B.1, Lemma 5.



(a)

(b)

Figure 5: It is sufficient to consider only horizontal, vertical and diagonal (along the isobelief curve) deviations. The blue curves depict potential isobelief curves. Red arrows are deviations.

The above result allows us to disentangle the effects of misreporting (1) one's payoff type – i.e., belief – and (2) the amount of evidence, as detailed below.

The traditional mechanism design with evidence literature mostly focuses on the case where (1) the agent is privately fully informed of their *payoff type*, and (2) evidence is not payoff-relevant to either player but plays a role only in restricting deviations (Green and Laffont (1986), Forges and Koessler (2005), Bull and Watson (2007*a*), Deneckere and Severinov (2008), Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), Sher and Vohra (2015), Celik (2006), KrÃ€hmer and Strausz (2024), Dasgupta, Krasikov and Lamba (2022),Vaidya (2023)).[9] A well-known idea in this literature, is that incentive constraints can be decomposed into two components: (1) a "soft" constraint, which prevents misreporting one's payoff type, with *no* directional restrictions, and (2), a "hard", *disclosure constraint*, with the directional restriction, that one is allowed to disclose only a subset of one's evidence (Deneckere and Severinov, 2008; Bull and Watson, 2007*b*; Vaidya, 2023). The idea behind this principle is the following. By the revelation principle appropriate to settings of mechanism design with evidence (Bull and Watson, 2007*b*; Deneckere and Severinov, 2008), in these settings, the design problem decomposes into a family of standard mechanism design problems (i.e., without verifiability), each conditional on an evidence level, and linked together by a disclosure constraint which says that each type must find it optimal to disclose all its evidence, *conditional on* reporting the payoff type truthfully.[10]

In contrast, *all* private information in our model is hard. The "decomposition" result described above shows that, in spite of that, our incentive constraints can be decomposed into "soft" and

[9] An exception is Dasgupta, Krasikov and Lamba (2022), where the agent learns purely from hard evidence, as in our setting.

[10] E.g., see Vaidya (2023) for a clear application of this principle to the setting of monopolistic selling with regulation.

"hard" components, exactly in the same way as in the more traditional setting described above. The diagonal incentive constraints are our *disclosure constraints*, in the language of the literature – this is the set of constraints for which the restrictions on the directions of deviation, matter, in the following sense. If each type was allowed to deviate along its isobelief line in *both* directions, but restricted to only leftward and downwards deviations horizontally and vertically respectively, we would effectively be back in Dasgupta (2024)'s setting, who considers essentially the same problem, but allows for only belief-based – as opposed to evidence-based – screening. In other words, the unidirectionality of only the diagonal incentive constraints have a bite in our setting, not the horizontal and vertical ones. In that sense, the horizontal and vertical constraints can be considered "soft".

4.3 Characterization of optimal mechanisms

We formally state the principal's problem. Let $v_R(r,l) := u(r,l)f(r,l)$ and $v_L(r,l) := u(l,r)f(l,r)$. The principal solves

$$\max_{a_R, a_L \in [0,1]^E} \int_E \left[v_R(r,l)a_R(r,l) + v_L(r,l)a_L(r,l) \right] drdl \tag{2}$$

subject to the IC constraints (1).

The optimal mechanism turns out to be a True-False test with an evidence requirement, as summarized in the following theorem.

Theorem 1. *There exists an optimal mechanism of the following form,*

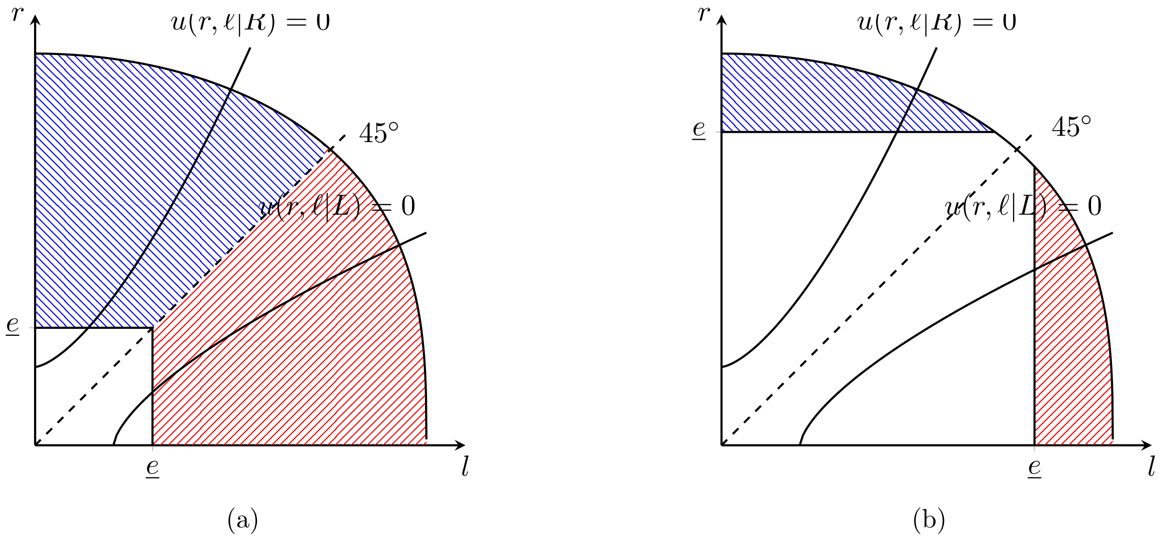
$$\begin{aligned} a_R(r,l) &= 1(r \geq \max\{\underline{e}, l\}), \\ a_L(r,l) &= 1(l \geq \max\{\underline{e}, r\}), \end{aligned} \tag{3}$$

where $\underline{e}$ is optimally chosen by

$$\underline{e} \in \arg\max_{e} \int\limits_{r \geq \max\{e,l\}} u(r,l)F(dr,dl). \tag{4}$$

Theorem 1 says that the optimal mechanism passes the agent if and only if he gets the correct answer and presents enough evidence, as illustrated in Fig. 2. $\underline{e}$ is the threshold amount of evidence required for the agent to be passed in any state. The principal's problem is then converted into a one-dimensional problem of choosing the optimal $\underline{e}$. Examples, for a general resource constraint $\phi(r,l) \leq 1$ and general preferences of the principal, satisfying Assumption 2, are depicted in the Figure below.

There are three features of the optimal mechanism. First, the optimal mechanism always passes some evidence types, provided that some evidence types are passed under the first best. This is due to the presence of verifiable evidence. Whenever some evidence types are passed under the first



(a)

(b)

Figure 6: Optimal mechanisms for a general (i.e., potentially non-linear) ϕ and $u(r,l)=0$ frontiers

best, assumption 3 implies that types with the maximum amount of correct evidence must have a positive value to the principal. The principal can pass only these types by setting a high enough threshold for correct evidence. If the agent's private information were unverifiable, it can be optimal to fail every type if the benefits from passing the good types do not justify the information rents to induce truth-telling.

Second, the threshold evidence for passing in each state depends only on the amount of evidence supporting this state. Intuitively, this is because of both the verifiability of evidence and the principal's preference for accuracy. Focus on state R and types with $r \geq l$. Suppose $\psi(l)$ is the threshold on r for passing if the agent has an amount l of evidence favoring L. $\psi(l)$ cannot slope upwards due to the verifiability of evidence. This is because if the principal passes some (r,l), then any $(r',l') > (r,l)$ should also be passed. Next, any downward sloping $\psi(l)$ cannot be optimal due to Assumption 3. This is because the principal can improve by turning some part of $\psi(l)$ into a horizontal line. To see this, first observe that for any $r_0 < \psi(0)$ there must exist some $l_0 \leq r_0$ with $r_0 \geq \psi(l_0)$ and $v_R(r_0,l_0) > 0$. Otherwise, $v_R(r,l) < 0$ for any $l \leq r < r_0$ due to Assumption 3, and the principal can improve by failing all types below r_0. We now change the threshold $\psi(l)$ to be $\psi'(l) := \min\{r_0,\psi(l)\}$ so that the principal now passes (r,l) with $r > \min\{r_0,\psi(l)\}$. The newly passed types all have $v_R(r,l) > 0$ again due to Assumption 3. Therefore, the optimal threshold $\psi(l)$ must not depend on l.

Third, an agent is passed only in the state where he has more supporting evidence. That is, type (r,l) with $r \geq l$ is never passed in state L, and vice versa for $l \geq r$. This is due to self-selection. Suppose instead (r,l) with $r \geq l$ is passed in state L but not R. He prefers to be passed only in state R because he believes that R is more likely. Since $r \geq l$ and the environment is symmetric, he must be able to conceal some l evidence to be passed in state R. Therefore, the optimal mechanism should pass types with $r \geq l$ only in state R and types with $l \geq r$ only in state L to incentivize full disclosure of evidence.

4.4 Proof Sketch

We briefly sketch the proof idea here. The complete proof is available in Appendix B. We use a duality approach to certify the optimality of (3) which may be of independent interest. At the core of the argument, we view the construction of dual multipliers as a transport problem, and apply Strassen (1965)'s theorem to prove the existence of such multipliers.

We illustrate the multiplier construction with a two-type example. Suppose there are only two types with an amount r of evidence favoring state R. Let them be (r, l) and (r, l') with $l' > l$. Consider the case where $r > \underline{e}$.[11] In this case, the mechanism in (3) passes both types in state R. Focus on the interesting case where $v_R(r, l) > 0, v_R(r, l') < 0$.[12] Without IC constraints, the principal only wants to pass (r, l) in state R. She has to pass both types due to the IC constraint from (r, l') to (r, l) which requires

$$a_R(r, l')\gamma(r, l') + a_L(r, l') \geq a_R(r, l)\gamma(r, l') + a_L(r, l).$$

We want to construct a multiplier λ on this IC constraint to certify the optimality of (3). To do that, λ has to make the virtual values positive for evidence types passed by the proposed mechanism, and negative for those failed,

$$\hat{v}_R(r, l') = v_R(r, l') + \gamma(r, l')\lambda \geq 0, \tag{5}$$

$$\hat{v}_L(r, l') = v_L(r, l') + \lambda \leq 0, \tag{6}$$

$$\hat{v}_R(r, l) = v_R(r, l) - \gamma(r, l')\lambda \geq 0, \tag{7}$$

$$\hat{v}_L(r, l) = v_L(r, l) - \lambda \leq 0. \tag{8}$$

The multiplier λ serves as a transport plan that specifies how virtual values are redistributed across evidence types. [13] It specifies how values of v_R and v_L are simultaneously transported from (r, l) to (r, l'). A transport plan must be measure-preserving in the sense that it only redistributes virtual values but does not create or destroy them. This is indeed the case according to (5)-(8). λ decreases the value of $v_R(r, l)$ (respectively, $v_L(r, l)$) and increases the value of $v_R(r, l')$ (respectively, $v_L(r, l')$) by the same amount.

[11] The case of $r < \underline{e}$ is similar.

[12] If $v_R(r, l) \geq 0$ and $v_R(r, l') \geq 0$, there is no need to construct any multiplier as the principal wants to pass them in state R. $v_R(r, l) < 0$ and $v_R(r, l') < 0$ cannot happen at the same time because otherwise the principal can raise $\underline{e}$ above r to fail both types in state R and obtain a strictly higher payoff, contradicting the optimality of $\underline{e}$. $v_R(r, l) > 0$ and $v_R(r, l') < 0$ cannot happen at the same time because this contradicts Assumption 3.1 that the principal prefers types with more correct evidence.

[13] Formally, a transport plan is a measure on the product space of the source and target domains that satisfies the given marginal constraints. In this example, the source is (r, l) and the target is (r, l'). The marginal constraints require that the values transported out of (r, l) must equal to the values transported into (r, l').

We now show that $\lambda = -\frac{v_R(r,l')}{\gamma(r,l')}$ satisfies all the above inequalities. The value of λ is picked so that (5) binds. (6) is satisfied because Assumption 3 implies that $v_R(r,l) - \gamma(r,l)v_L(r,l) \geq 0$ for any (r,l).[14] (7) holds due to $v_R(r,l) + v_R(r,l') \geq 0$, which comes from the optimality of $\underline{e}$. To see this, if $v_R(r,l) + v_R(r,l') < 0$, the principal can raise $\underline{e}$ above r to fail both types and obtain a strictly higher payoff. (8) holds because $v_L(r,l) \leq 0$.

More generally, we can always construct multipliers to certify the optimality of (3) for $r \geq \underline{e}$ if $v_R(r,l)$, viewed as a measure, has enough positive mass to be shifted to the northeast to fill up its negative mass. Strassen's theorem makes this point formal. It says that a sufficient condition for such multipliers to exist is that $v_R^+(r,l) := \max\{v_R(r,l), 0\}$ is first order stochastically dominated by $v_R^-(r,l) := \max\{-v_R(r,l), 0\}$. This sufficient condition is always satisfied due to the optimality of $\underline{e}$ given by (4). The proof for $r \leq \underline{e}$ is similar and the details are provided in Appendix B.

5 Microfounded Model

In this section we present a microfounded model which leads to the reduced form with our desired features, discussed in the main model section. The microfounded model allows us to run some comparative statics on the previously introduced evidence threshold $\underline{e}$ of our optimal class of mechanisms.

Recall the setting of the example from Section 3. Essentially, in this section we generalize it. First, we do not require $u_B \geq 1$. Second, we allow for general prior distributions over the Good and Bad types in the population – let the prior proportion of the Good type be g. Finally, we generalize the joint distribution among the state, quality and evidence, as follows. There is a distribution over evidence in each state for each quality-type of agent, which is assumed to have full support on the evidence space E and a density. Like before, let $f_q(r,l)$ denote the density of evidence type (r,l) in state R for quality-type $q \in \{G, B\}$, which means, by symmetry, $f_q(l,r)$ is that in state L.

We impose the following assumptions on f_B and f_G:

Assumption 6. *1. Each of f_G and f_B is strictly increasing in r and strictly decreasing in l.*

2. $\frac{\frac{\partial f_G(r,l)}{\partial r}}{f_G(r,l)} \geqslant \frac{\frac{\partial f_B(r,l)}{\partial r}}{f_B(r,l)}$, $\frac{\left|\frac{\partial f_G(r,l)}{\partial l}\right|}{f_G(r,l)} \geqslant \frac{\left|\frac{\partial f_B(r,l)}{\partial l}\right|}{f_B(r,l)}$

3. $\frac{\frac{\partial f_G(r,l)}{\partial r}}{\frac{\partial f_B(r,l)}{\partial r}} \geqslant \frac{\left|\frac{\partial f_G(r,l)}{\partial l}\right|}{\left|\frac{\partial f_B(r,l)}{\partial l}\right|}$

The second assumption essentially captures the feature that the Good type's evidence distribution is more sensitive to *both* the correct and wrong evidence. Recall that $f_G(r,l)$ and $f_B(r,l)$ are the densities in state R, so r is the *correct* evidence. Keeping that in mind, the third assumption tells us that the relative sensitivity of the good vis-a-vis the bad quality, to the *correct* evidence is higher than that for the wrong evidence.

[14] See Lemma 4 in Appendix B.

In the appendix we show that Assumptions 6 are sufficient for Assumptions 3.

$u(r,l)$, as defined in our model section is the expected utility of the principal from accepting the evidence type (r,l) in state R. Here, it is the interim expected quality of the agent. Using Bayes rule, some algebra gives us:

$$u(r,l) = \frac{f_G(r,l) - \alpha f_B(r,l)}{f_G(r,l) + \xi f_B(r,l)} \qquad \text{(microfoundation - } u(r,l)\text{)}$$

where $\alpha := u_B \frac{1-g}{g}$ and $\xi := \frac{1-g}{g}$.

5.1 Comparative Statics

In this subsection we show that the evidence threshold $\underline{e}$ of the optimal mechanism is monotonically increasing with the principal's *quality sensitivity*.

In particular, we show it monotonically increases with α, defined above. α the single parameter which captures all relevant aspects of the principal's preference primitives, in our microfounded model. It captures her *quality sensitivity* – clearly, it increases both as (1) the principal's loss from accepting the bad quality agent in increases, and (2) as Good quality agents become increasingly rare in the population. In the next subsection we show that the evidence threshold $\underline{e}$ of the optimal mechanism is monotonic with α.

Proposition 2. *$\underline{e}$ is monotonically increasing in the principal's quality sensitivity α.*

In particular, with reference to figures 6, the $u(r,l|R) = 0$ and $u(r,l|L) = 0$ curves shift outwards as α increases. This is because, recall from equation (microfoundation - $u(r,l)$), that,

$$u(r,l) = 0 \Leftrightarrow \frac{f_G(r,l)}{f_B(r,l)} = \alpha$$

Consequently, so does the optimal choice of $\underline{e}$.

Intuitively, a higher quality sensitivity on the part of the principal – her higher relative loss from passing the bad quality agent (a high u_B), *or* good quality agents being hard to find in the population (a high g) – both push her to make the test "harder", increasing the evidence threshold.

6 An extension

We now explore how the optimal mechanisms change if the principal may prefer to pass some types in both states. That is, drop Assumption 4 from the baseline model. It turns out that, under a regularity condition, the optimal mechanism now features two evidence threshold, namely, $\underline{e}$ and $\bar{e}$. Similar to Theorem 1, the optimal mechanism fails the agent with certainty if he has too little evidence (less than $\underline{e}$). The agent is given a True-False test and passed if he gives the correct answer

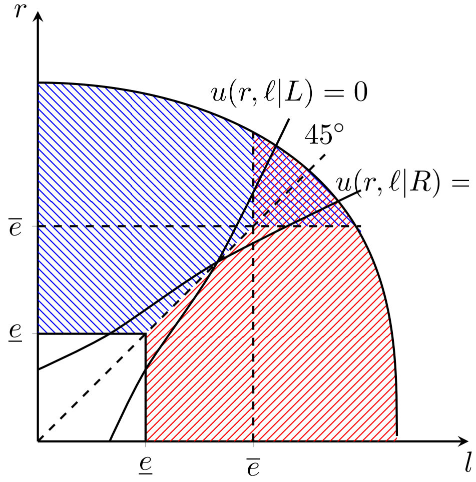


Figure 7: Optimal Mechanism with 4 Regions

if he possesses an intermediate level of evidence (at least $\underline{e}$ for any state). The optimal mechanism now includes an additional case. The agent is passed regardless of his answer if he presents enough evidence for both states (at least $\bar{e}$ for both states).

First, we present and interpret the regularlity condition we need. We state it in terms of state R. By symmetry, it also implies its version when r and l are interchanged and R is changed to L.

Assumption 7. *The positive part of the principal's (probability-weighted) expected payoff from accepting all evidence types in* $\{\min\{r,l\} \geq \tilde{e}, r \geq l\}$ *in state* L*, is quasiconcave in* $\tilde{e}$*. Mathematically,*

$$\left(\int\limits_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right)^{+}$$

is quasiconcave in $\tilde{e}$*.*

The interpretation of the above assumption is as follows. The passing rule $\{\min\{r,l\} \geq \tilde{e}, r \geq l\}$ captures passing specifically those who have a high enough evidence of *both* kinds ($\min\{r,l\} \geq \tilde{e}$), but hold the *wrong* belief ($r \geq l \implies$ these types wrongly believe that state R is more likely, but the actual state is L), i.e., score low on accuracy. Therefore under this passing rule, there is an evidence/accuracy trade-off on the part of the principal, i.e. this function is non monotonic. Here, we assume it is single-peaked.

Clearly, the above assumption is trivially satisfied under our original assumption in the main model, Assumption 4, because $\left(\int\limits_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right)^{+} = 0$ in that case.

Now we are ready to state our main result, under this assumption. Fig. 7 visualizes the optimal mechanism under Assumption 7.

Theorem 2. *Under Assumption 3 and the regularity condition 7, there exists an optimal mechanism of the following form,*

$$\begin{aligned} a_R(r,l) &= 1(r \geq \min\{\max\{\underline{e}, l\}, \bar{e}\}), \\ a_L(r,l) &= 1(l \geq \min\{\max\{\underline{e}, r\}, \bar{e}\}), \end{aligned} \tag{9}$$

where $\underline{e}$ and $\bar{e}$ are optimally chosen by

$$\underline{e}, \bar{e} \in \arg\max_{e_1, e_2} \int_{\max\{e_1, l\} \leq r \leq e_2} u(r,l) F(dr, dl). \tag{10}$$

Relating this to Assumption 7, $\bar{e}$ is the optimal $\tilde{e}$ for $\left(\int_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right)^+$.

The significance of Theorem 2 is that it shows that the acceptance decision is purely on the basis of evidence whenever there is either too little ($\max\{r,l\} < \underline{e}$) or too much ($\min\{r,l\} > \bar{e}$). For intermediate levels of evidence, the agent is accepted if and only if they give the correct answer. This matches with evaluation schemes we observe in reality where students are sometimes rewarded if they show enough working, or make enough arguments, even if they cannot get the ultimate answer correct.

The proof for Theorem 2 uses similar "transportation of mass" techniques described in the proof sketch of our main characterization.

7 Conclusion

We consider the setting where a test-taker (agent) is screened on the basis of his knowledge of a binary state, and needs to provide justifications for his answers/"show his work". He learns *only* through verifiable evidence, of two contradictory kinds. The test-designer (principal) has a preference for both the *accuracy* of his knowledge (modeled as the precision of his belief about the state), and the *amount* of evidence he can show in support of his answers. Due to the verifiable nature evidence, the agent can only hide part of his evidence, but cannot manufacture any, lending our incentive constraints a "unidirectional" nature. We delineate how the interaction of verifiability and the principal's preference for both evidence and accuracy leads to the novel *evidence/accuracy trade-off* on the part of the agent. This leads to his desired deviations being type-dependent, in a departure from the mechanism-design-with-evidence literature so far. This is the reason unidirectionality has bite in our setting, unlike many of the settings previously considered (Celik, 2006; KrÃ€hmer and Strausz, 2024).

When parameters are such that there is no evidence type which would be accepted by the principal in both states under her first-best mechanism, we show that the optimal mechanism takes the form of *True-False with an evidence threshold.* This means, the agent is accepted if and only if he correctly predicts the state, but conditional on showing a minimum level of supporting evidence.

When there are types the principal would ideally want to accept in both states, under a regularity condition, we show that the optimal mechanism exhibits an additional upper threshold of evidence: In particular, the optimal test passes (respectively, fails) the agent regardless of his answer, if the amount of evidence provided is sufficiently high (respectively, sufficiently low), and passes him if and only if his answer is correct, when the amount of evidence provided is intermediate.

Taken together, these results provide rationalization for commonly observed test structures, where students, job interview candidates etc. are rewarded for the correctness of their answers, only if they provide a minimum amount of reasoning/justifications for their answers, and conversely, they may also be rewarded if they show enough reasoning, but are unable to arrive precisely at the correct answer.

References

Abernethy, Jacob D, and Rafael M Frongillo. 2012. "A characterization of scoring rules for linear properties." 27–1, JMLR Workshop and Conference Proceedings.

Adams, William J., and Janet L. Yellen. 1976. "Commodity Bundling and the Burden of Monopoly." *The Quarterly Journal of Economics*, 90(3): 475–498.

Armstrong, Mark. 1996. "Multiproduct Nonlinear Pricing." *Econometrica*, 64(1): 51–75.

Ben-Porath, Elchanan, and Bart Lipman. 2012. "Implementation with Partial Provability." *Journal of Economic Theory*, 147(5): 1689–1724.

Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman. 2014. "Optimal allocation with costly verification." *American Economic Review*, 104(12): 3779–3813.

Bull, Clifford, and Joel Watson. 2007*a*. "Hard Evidence and Mechanism Design." *Games and Economic Behavior*, 58(1): 75–93.

Bull, Jesse, and Joel Watson. 2007*b*. "Hard evidence and mechanism design." *Games and Economic Behavior*, 58(1): 75–93.

Cai, Yang, Nikhil R. Devanur, and S. Matthew Weinberg. 2019. "A Duality-Based Unified Approach to Bayesian Mechanism Design." *SIAM Journal on Computing*, 48(0): STOC16–160.

Carroll, Gabriel. 2017. "Robustness and Linear Contracts." *American Economic Review*, 107(2): 59–86.

Carroll, Gabriel, and Georgy Egorov. 2019. "Strategic communication with minimal verification." *Econometrica*, 87(6): 1867–1892.

Celik, Gorkem. 2006. "Mechanism Design with Weaker Incentive Compatibility Constraints." *Games and Economic Behavior*, 56(1): 37–44.

Chambers, Christopher P, and Nicolas S Lambert. 2021. “Dynamic belief elicitation.” *Econometrica*, 89(1): 375–414.

Dasgupta, Sulagna. 2024. “Optimal Test Design for Knowledge-based Screening.” *Available at SSRN 4403119.*

Dasgupta, Sulagna, Ilia Krasikov, and Rohit Lamba. 2022. “Hard Information Design.” SSRN working paper.

Daskalakis, Constantinos, Alan Deckelbaum, and Christos Tzamos. 2017. “Strong Duality for a Multiple-Good Monopolist.” *Econometrica*, 85(3): 735–767.

Deb, Rahul, Mallesh M Pai, and Maher Said. 2018. “Evaluating strategic forecasters.” *American Economic Review*, 108(10): 3057–3103.

Deb, Rahul, Mallesh M Pai, and Maher Said. 2023. *Indirect Persuasion.* Centre for Economic Policy Research.

Deneckere, Raymond J., and Sergei Severinov. 2008. “Mechanism Design with Partially Verifiable Information.” *Games and Economic Behavior*, 64(2): 487–513.

Dziuda, Wioletta. 2011. “Strategic argumentation.” *Journal of Economic Theory*, 146(4): 1362–1397.

Forges, FranÃ§oise, and FrÃ©dÃ©ric Koessler. 2005. “Communication Equilibria with Partially Verifiable Types.” *Journal of Mathematical Economics*, 41(7): 793–811.

Glazer, Jacob, and Ariel Rubinstein. 2004. “On optimal rules of persuasion.” *Econometrica*, 72(6): 1715–1736.

Green, Jerry R., and Jean-Jacques Laffont. 1986. “Partially Verifiable Information and Mechanism Design.” *The Review of Economic Studies*, 53(3): 447–456.

Haghpanah, Nima, and Jason Hartline. 2021. “When Is Pure Bundling Optimal?” *The Review of Economic Studies*, 88(3): 1127–1156.

Hancart, Nathan. 2022. “Designing the Optimal Menu of Tests.”

Harbaugh, Rick, and Eric Rasmusen. 2018. “Coarse grades: Informing the public by withholding information.” *American Economic Journal: Microeconomics*, 10(1): 210–235.

Kartik, Navin, and Olivier Tercieux. 2012. “Implementation with Evidence.” *Theoretical Economics*, 7(2): 323–355.

KrÃ€hmer, Daniel, and Roland Strausz. 2024. “Unidirectional Incentive Compatibility.” CRC TR 224 Discussion Paper No. 524.

Lambert, Nicolas S. 2011. “Elicitation and evaluation of statistical forecasts.” *Preprint.*

Li, Yingkai, and Jonathan Libgober. 2023. "Optimal Scoring for Dynamic Information Acquisition." *arXiv preprint arXiv:2310.19147*.

Li, Yingkai, Jason D Hartline, Liren Shan, and Yifan Wu. 2022. "Optimization of scoring rules." 988–989.

Manelli, Alejandro M., and Daniel R. Vincent. 2007. "Multidimensional Mechanism Design: Revenue Maximization and the Multiple-Good Monopoly." *Journal of Economic Theory*, 137(1): 153–185.

Marinovic, Iván, Marco Ottaviani, and Peter Sorensen. 2013. "Forecastersâ objectives and strategies." In *Handbook of economic forecasting*. Vol. 2, 690–720. Elsevier.

McAfee, R. Preston, and John McMillan. 1988. "Multidimensional Incentive Compatibility and Mechanism Design." *Journal of Economic Theory*, 46(2): 335–354.

McAfee, R. Preston, John McMillan, and Michael D. Whinston. 1989. "Multiproduct Monopoly, Commodity Bundling, and Correlation of Values." *The Quarterly Journal of Economics*, 104(2): 371–383.

McCarthy, John. 1956. "Measures of the value of information." *Proceedings of the National Academy of Sciences*, 42(9): 654–655.

Müller, Alfred, and Dietrich Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks.* Chichester, England:John Wiley & Sons.

Osband, Kent, and Stefan Reichelstein. 1985. "Information-eliciting compensation schemes." *Journal of Public Economics*, 27(1): 107–115.

Ottaviani, Marco, and Peter Norman Sørensen. 2006*a*. "Professional advice." *Journal of Economic Theory*, 126(1): 120–142.

Ottaviani, Marco, and Peter Norman Sørensen. 2006*b*. "Reputational cheap talk." *The Rand journal of economics*, 37(1): 155–175.

Pavlov, Gregory. 2011. "Optimal Mechanism for Selling Two Goods." *The BE Journal of Theoretical Economics*, 11(1): 1–35.

Rochet, Jean-Charles, and Philippe ChonÃ©. 1998. "Ironing, Sweeping, and Multidimensional Screening." *Econometrica*, 66(4): 783–826.

Rosar, Frank. 2017. "Test design under voluntary participation." *Games and Economic Behavior*, 104: 632–655.

Sher, Itai, and Rakesh Vohra. 2015. "Price Discrimination through Communication." *Theoretical Economics*, 10(2): 597–648.

Stigler, George J. 1963. "United States v. Loew's Inc.: A Note on Block Booking." *The Supreme Court Review*, 1963: 152–157.

Strassen, Volker. 1965. "The existence of probability measures with given marginals." *The Annals of Mathematical Statistics*, 36(2): 423–439.

Vaidya, Udayan. 2023. "Regulating Disclosure: The Value of Discretion." [Online; accessed June-18-2023].

Weksler, Ran, and Boaz Zik. 2022. "Informative tests in signaling environments." *Theoretical Economics*, 17(3): 977–1006.

Yang, Frank. 2022. "Costly Multidimensional Screening."

Yang, Frank. 2023. "Nested Bundling." Previously titled "The Simple Economics of Optimal Bundling".

Yang, Frank, Alexander Haberman, and Ravi Jagadeesan. 2025. "Multidimensional Screening with Returns." Working paper, last updated January 2025.

Yang, Frank, Matthew Gentzkow, Jesse M. Shapiro, and Ali Yurukoglu. 2024. "Pricing Power in Advertising Markets: Theory and Evidence." *American Economic Review*, 114(forthcoming).

Yang, Frank, Piotr Dworczak, and Mohammad Akbarpour. 2023. "Comparison of Screening Devices." Revise & Resubmit, Journal of Political Economy.

A Related Literature

This paper mainly contributes to three distinct strands of literature – screening knowledge, mechanism design with evidence, and multi-dimensional mechanism design. Less directly, it also contributes to the literature on mechanism design with verificaiton and that on test design.

First, this paper builds on the literature on screening agents on the basis of their "knowledge", modeled as beliefs, as in (Dasgupta, 2024) and Deb, Pai and Said (2023). Dasgupta (2024) is related most closely to our work, who considers the same question – optimal mechanisms to evaluate a test-taker on the basis of their knowledge – except all private information is unverifiable in her model. Deb, Pai and Said (2023) consider a joint screening-and-persuasion problem and find a similar characterization of the class of optimal mechanisms, as in Dasgupta (2024). This literature, in turn, builds on the literatures on scoring rules (McCarthy (1956), Osband and Reichelstein (1985), Lambert (2011), Abernethy and Frongillo (2012), Li et al. (2022), Li and Libgober (2023)) and evaluation of forecasters (Deb, Pai and Said (2018), Chambers and Lambert (2021), Ottaviani and Sørensen (2006*a*), Ottaviani and Sørensen (2006*b*); also see Marinovic, Ottaviani and Sorensen (2013) for a survey).

This paper contributes to the literature on mechanism design with evidence (Green and Laffont (1986), Forges and Koessler (2005), Bull and Watson (2007*a*), Deneckere and Severinov (2008), Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), etc.). Within this literature, this paper relates most closely to Sher and Vohra (2015), Celik (2006), KrÃ€hmer and Strausz (2024), Dasgupta, Krasikov and Lamba (2022) and Vaidya (2023). Sher and Vohra (2015), Dasgupta, Krasikov and Lamba (2022) and Vaidya (2023), all have the classic monopolistic screening problem at the core of their models. Dasgupta, Krasikov and Lamba (2022) allows flexible acquisition of "all or nothing" evidence by the agent (buyer), in addition. While Vaidya (2023) permits arbitrary correlation between the agent's valuation and his evidence, he also allows the agent to only present or not present his evidence to the principal (seller). In contrast, we allow the agent to choose the *amount* of evidence he presents, which is bounded upwards by the amount he has. This restricts the "directions" of misreporting by the agent, like in Celik (2006) and KrÃ€hmer and Strausz (2024). Sher and Vohra (2015) allows for an evidence structure which can be thought of as a generalization of all of the above cases, where each type can mimic some of the other types but not necessarily all of them. None of these papers consider multi-dimensional evidence. Our two-dimensional evidence structure is similar to Dziuda (2011)'s, though her problem is entirely different.

Our most fundamental point of departure from the above literature is that in much of it, the principal and the agent's payoff relevant information is the same – e.g., a buyer's valuation, in case of the classic selling problem – and evidence is used only to support claims regarding that information. To the best of our knowledge, ours is the first paper to consider the case where the *amount* of evidence has direct relevance to the principal, beyond its indirect relevance through informing the agent about his payoffs.

The agent's evidence is two-dimensional in this paper. This leads to the usual complications of multi-dimensional screening, and connects it to the relevant, vast literature (Stigler (1963); Adams and Yellen (1976); McAfee, McMillan and Whinston (1989); Armstrong (1996); Rochet and ChonÃ© (1998); Carroll (2017); McAfee and McMillan (1988); Manelli and Vincent (2007); Pavlov (2011); Daskalakis, Deckelbaum and Tzamos (2017); Yang (2023, 2022); Yang, Dworczak and Akbarpour (2023); Yang, Haberman and Jagadeesan (2025); Yang et al. (2024)). Unlike this paper, much of the aforementioned literature studies the multi-dimensional screening problem in a multi-good selling setting. Hence, our work is related to this literature primarily in terms of technique. Our proof uses a duality approach to certify optimality, leveraging a novel construction of virtual values. In this regard, within this literature, our work is closest to Haghpanah and Hartline (2021), which in turn builds on Carroll (2017) and Cai, Devanur and Weinberg (2019).

At a broader level, this paper also relates to the literature on mechanism design with verification but without transfers. Like in our paper, in this literature the instrument for eliciting private information from strategic agents is – not monetary incentives, but – information obtainable by the principal. Closest to our work within this literature are Glazer and Rubinstein (2004) and Carroll and Egorov (2019). In their models a principal accepts or rejects an agent based on limited verification of his claimed “quality”. Also related, although less closely, is Ben-Porath, Dekel and Lipman (2014) which features a similar multi-agent model, but with exact verification at a cost.

In terms of our question – though not very closely in terms of our model or methods – this work also relates to the literature on how a receiver of information (in our case, the principal) designs a test of some unobservable quality of a strategic sender (the agent) (Rosar (2017); Harbaugh and Rasmusen (2018); Weksler and Zik (2022); Hancart (2022)). Much of this literature leverages information design tools to characterize optimal tests in various environments. A common finding of this literature is that more informative tests are not always better, due to the strategic incentives such tests create for the agent. In particular, similar to our paper, some of this literature finds *coarse* tests arising at the optimum (Rosar (2017), Harbaugh and Rasmusen (2018)).

B Omitted Proofs

B.1 Preliminaries

We now provide some preliminary results for our proofs.

Lemma 1. *Assumption 2 and assumption 3 are equivalent.*

Proof. For the sake of brevity, we provide the proof for the case where u is differentiable in both arguments. The proof is easily adjustable for the case where the type space is discrete.

Let $p_0 = p(r, l)$, i.e., $l = p^{-1}(p_0; r)$, so $\tilde{u}(p_0, r) = u(r, l)$.

Assumption 3 ⇒Assumption 2 We first show that, $\frac{\partial \tilde{u}}{\partial p_0} \geq 0$. $\frac{\partial \tilde{u}}{\partial p_0} = \frac{\partial u}{\partial l} \cdot \frac{\partial l}{\partial p_0} = \frac{\partial u}{\partial l} \cdot \frac{1}{\frac{\partial p(r,l)}{\partial l}}$.

$$\therefore \frac{\partial \tilde{u}}{\partial p_0} = \frac{1}{\frac{\partial p(r,l)}{\partial l}} \cdot \frac{\partial u(r, l)}{\partial l} \tag{11}$$

By Assumption 3.1, $\frac{\partial u(r,l)}{\partial l} < 0$. By Assumption 1, $\frac{\partial p(r,l)}{\partial l}$, hence $\frac{\partial \tilde{u}}{\partial p_0} > 0$.

Next we show that, $\frac{\partial \tilde{u}}{\partial r} \geq 0$.

We have, $\frac{\partial \tilde{u}}{\partial r} = \frac{\partial u}{\partial r} + \frac{\partial u}{\partial l} \cdot \frac{\partial l}{\partial r}$, where $\frac{\partial l}{\partial r}$ is calculated along the curve $p(r, l) = \text{constant} = p_0$. $\therefore \ \frac{\partial l}{\partial r} = -\frac{\frac{\partial p(r,l)}{\partial r}}{\frac{\partial p(r,l)}{\partial l}}$.

$$\therefore \frac{\partial \tilde{u}}{\partial r} = \frac{\partial u}{\partial r} - \frac{\partial u}{\partial l} \cdot \frac{\frac{\partial p(r,l)}{\partial r}}{\frac{\partial p(r,l)}{\partial l}}, \tag{12}$$

According to Assumption 3.2, $\frac{du}{dr} \geq 0$ along a curve along which $p(r, l) = \text{constant}$, i.e.,

$$\left.\frac{\partial u}{\partial r}\right|_{p(r,l)=\text{constant}} \geq 0,$$

$$\text{i.e.,} \frac{\partial u}{\partial r} + \frac{\partial u}{\partial l} \cdot \left.\frac{\partial l}{\partial r}\right|_{p(r,l)=\text{constant}} = \frac{\partial u}{\partial r} - \frac{\partial u}{\partial l} \cdot \frac{\frac{\partial p(r,l)}{\partial r}}{\frac{\partial p(r,l)}{\partial l}} \geq 0.$$

By (12), this establishes $\frac{\partial \tilde{u}}{\partial r} \geq 0$.

The two parts taken together establishes Assumption 3.

Assumption 2 ⇒Assumption 3 By Assumption 1 and (11), $\frac{\partial \tilde{u}}{\partial p_0} > 0 \implies \frac{\partial u(r,l)}{\partial l} < 0$. By $\frac{\partial u(r,l)}{\partial l} < 0$ and Assumption 1, the term after the minus sign in (12) is positive. Therefore $\frac{\partial \tilde{u}}{\partial r} > 0 \implies \frac{\partial u}{\partial r} > \frac{\partial u}{\partial l} \cdot \frac{\frac{\partial p(r,l)}{\partial r}}{\frac{\partial p(r,l)}{\partial l}} > 0$. This shows that Assumptions 2.1-2 taken together imply Assumption 3.1. Assumption 2.2 is clearly equivalent to Assumption 3.2. This completes the proof. □

Lemma 2. *Let $I_u^0(l) := u^{-1}(0;l)$ be the principal's iso-value curve in state R, and $I_p^{p_0}(l) := p^{-1}(p_0;l)$ be the agent's iso-belief curve. The iso-belief curve is always steeper than the iso-value curve, formally, for any p_0 and l,*

$$\frac{\partial I_p^{p_0}(l)}{\partial l} > \frac{\partial I_u^0(l)}{\partial l}.$$

Proof. This uses the same idea as Lemma 1. Starting from any (r,l) with $u(r,l) = 0$, increase the amount of evidence alone the iso-belief line to $(r',l') > (r,l)$, assumption 2.2 implies that $u(r',l') > 0$. To move back to the iso-value curve $I_p^{p_0}(l)$, we have to increase l or decrease r due to assumption 2.1. Therefore, the iso-belief curve is always steeper than the iso-value curve. □

Lemma 3. *$v_R(l,r) \leq 0, v_L(r,l) \leq 0, \forall r \geq l$.*

Proof. Take any $(r,l) \in E$ with $r \geq l$. It suffices to show that $u(l,r) \leq 0$. Suppose instead $u(l,r) > 0$. Assumption 3.1 implies that $u(r,r) > 0$. But this directly contradicts Assumption 3.3. □

Lemma 4. *$v_R(r,l) - \gamma(r,l)v_L(r,l) \geq 0, \forall r \geq l$.*

Proof. Take any $(r,l) \in E$ with $r \geq l$. To show that $v_R(r,l) - \gamma(r,l)v_L(r,l) \geq 0$, we plug in the definitions of v_R, v_L and γ.

$$u(r,l)f(r,l) \geq \frac{\Pr(R \mid r,l)}{\Pr(L \mid r,l)} u(l,r)f(l,r) \tag{13}$$

$$\Leftrightarrow u(r,l)f(r,l) \geq \frac{f(r,l)}{r(l,r)} u(l,r)f(l,r) \tag{14}$$

$$\Leftrightarrow u(r,l) \geq u(l,r) \tag{15}$$

where we use Assumption 3.1 to see that

$$u(r,l) \geq u(l,l) \geq u(l,r).$$

□

Lemma 5 (Horizontal, vertical and local diagonal IC's are sufficient)**.** *For each evidence type* (r,l)*, it is without loss to ignore all IC's other than (1)* $(r,l) \to (r,l'), l' < l, (2)(r,l) \to (r',l), r' < r$ *and (3)* $(r,l) \to (r',l')$ *where* $p(r,l) = p(r',l'), r' < r, l' < l$*. Moreover, it is sufficient to consider only local IC's of the last nature.*

Proof of Lemma 5. Fix any optimal solution to the principal's problem. Suppose $(r,l) \to (r',l')$ binds where $r' < r, l', l$ and $p(r,l) \neq p(r',l')$. The line joining (r,l) and (r',l') is either strictly between the isobelief line and the line joining (r,l) and (r,l') or strictly between the isobelief line and the line joining (r,l) and (r',l), as shown in Fig. 8 below.

Suppose $A \to C$ binds. Which means A is indifferent between its own and C's allocation. That means B is also indifferent between A's and C's allocation, because A and B have the same preferences (belief). But B can deviate to C, so B must prefer its own allocation to C's. If B strictly prefers its own allocation to C's, it means B strictly prefers its own allocation to A's as well, because B is indifferent between A and C. That means A also prefers B's allocation strictly to its own, because, again, A and B are on the same isobelief line. This is a violation of the constraint $A \to B$, i.e. a contradiction. Hence, B must be indifferent among all three allocations, A, B, C, hence so must be A. That is, $A \to B$ binds and $B \to C$ binds. Clearly, the converse is also true – if $A \to B$ binds and $B \to C$ binds, $A \to C$ binds. Hence, any binding IC can be decomposed into binding IC's **along** and **across** isobelief lines – AC binds if and only if AB and BC bind.

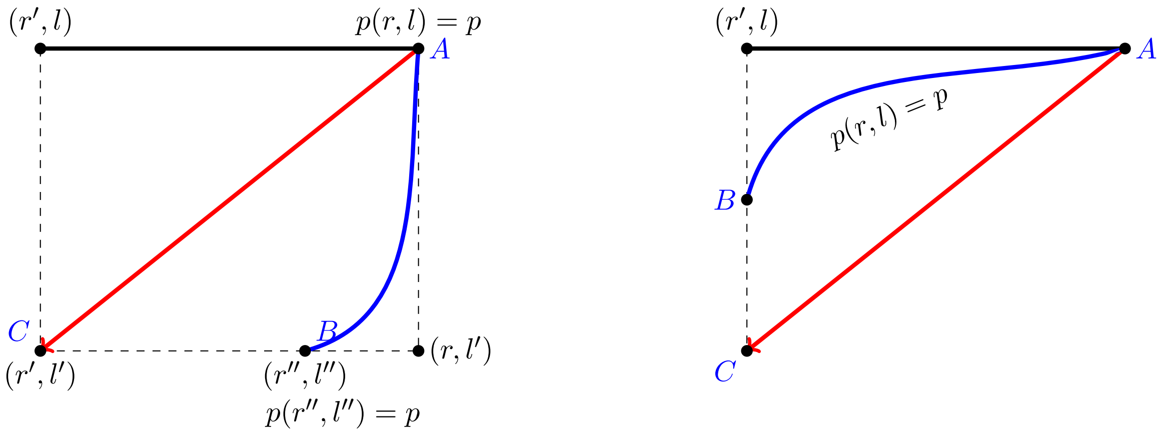


Figure 8: It is sufficient to consider only horizontal, vertical and diagonal (along the isobelief curve) deviations. The blue curves depict potential isobelief curves. Red arrows are deviations.

Let $U(r,l)$ denote the truth-telling utility of evidence type (r,l) under an IC mechanism. Let

$r > r' > r'', l > l' > l''$ such that $p(r,l) = p(r',l') = p(r'',l'')$.

By IC's $(r,l) \to (r',l')$ and $(r',l') \to (r'',l'')$ we need,

$$U(r,l) \geq U(r',l') \geq U(r'',l'') \tag{Diagonal IC}$$

the IC $(r,l) \to (r'',l'')$ binds if and only if, $U(r,l) = U(r'',l'')$. By (Diagonal IC), this means, $U(r,l) = U(r',l') = U(r'',l'')$. This shows that a non-local diagonal IC binds if and only if all local diagonal IC's in between them bind.

B.2 Strassen's Theorem and First Order Stochastic Dominance

We present a version of Strassen (1965)'s theorem and a characterization of first order stochastic dominance for measures over $\mathbb{R}^2$ (adapted from Theorem 3.3.4 of Müller and Stoyan (2002)). They will be useful for our proof of Theorem 1.

Theorem 3 (Strassen 1965)**.** *Let μ and μ' be probability measures on $\mathbb{R}^n$. Suppose $\mu \geq_{FOSD} \mu'$, then there exists a probability measure $\hat{\mu}$ on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals μ and μ' such that $\hat{\mu}(M) = 1$, where $M := \{(x, x') : x, x' \in \mathbb{R}^n, x \leq x'\}$.*

Before stating the second result, we introduce the following definition. Say that a set $S \in \mathbb{R}^n$ is a lower set if the indicator function 1_S is decreasing. Hence a set S is lower if and only if $x \in S$ and $x \geq y$ imply $y \in S$.

Theorem 4 (Adapted from Müller and Stoyan 2002)**.** *Let μ and μ' be probability measures on $\mathbb{R}^n$. $\mu \geq_{FOSD} \mu'$ if and only if $\mu(S) \geq \mu'(S)$ for any lower set S.*

B.3 Proof of Theorem 1

We now prove Theorem 1. We do so in several steps. First, due to symmetry, it suffices to consider problem (2) with half of the types and half of the constraints. Second, we incorporate the IC constraints with multipliers. Next, we observe that the multipliers serve as a transport plan that shifts values of v_R and v_L. We then apply Strassen (1965)'s theorem to prove the existence of such multipliers as a transport plan.

Invoking Symmetry It suffices to certify the optimality of (21) in the following relaxed problem with only half of the evidence type space $E_2 := \{(r,l) \in E : r \geq l\}$ and the IC constraints within E_2,

$$\max_{a_R, a_L \in [0,1]^{E_2}} \int_{E_2} [v_R(r,l)a_R(r,l) + v_L(r,l)a_L(r,l)]\, drdl \tag{16}$$
$$\text{s.t. } a_R(r,l)\gamma(r,l) + a_L(r,l) \geq a_R(r',l')\gamma(r,l) + a_L(r',l'), \forall (r,l), \big(r',l'\big) \in E_2, \big(r',l'\big) \leq (r,l)$$

Note that we divide the IC constraints by $p(l,r)$ on both sides.

To see this, note that (16) changes the original problem (2) in two ways. First, it relaxes the IC constraints (1) by ignoring deviations from any (r,l) with $r > l$ to any (r',l') with $r' < l'$ and vice versa. Second, it invokes symmetry to focus on half of the types in E_2 because principal's value from the other half is exactly the same. As a result, if (3) is optimal in the relaxed problem (16), it should also be optimal in the original problem (2) because it is feasible under the constraints of the original problem.

Constructing Virtual Values We attack the problem by constructing a set of multipliers that certifies the optimality of the mechanism proposed in (3). Let $\Lambda_{r,l}^{r',l'}$ be the multiplier on the IC constraint of type (r',l') misreporting as type $(r,l) \leq (r',l')$. Mathematically, $\Lambda_{r,l}^{r',l'}$ is a positive measure on $E_2 \times E_2 := \{(r',l',r,l) : (r',l'),(r,l) \in E_2\}$. $\Lambda_{r,l}^{r',l'}$ can only place mass on deviations from (r',l') to some $(r,l) \leq (r',l')$. Formally, this requires

$$\Lambda(E_2 \setminus D) = 0 \tag{17}$$

where $D := \{(r',l',r,l) : (r',l'),(r,l) \in E_2, (r',l') \geq (r,l)\}$.

Problem (16) becomes the following unconstrained problem,

$$\max_{a_R,a_L \in [0,1]^{E_2}} \int_{(r,l)\in E_2} [\hat{v}_R(r,l)a_R(r,l) + \hat{v}_L(r,l)a_L(l,r)]\, \mathrm{d}r\mathrm{d}l \tag{18}$$

where the virtual values $\hat{v}_R$ and $\hat{v}_L$ are defined by

$$\begin{aligned}\hat{v}_R(r,l) &:= v_R(r,l) + \int_{(r',l')\leq(r,l)} \gamma(r,l)\Lambda_{\mathrm{d}r',\mathrm{d}l'}^{r,l} - \int_{(r',l')\geq(r,l)} \gamma(r',l')\Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} &(19)\\ &= v_R(r,l) + \gamma(r,l)\Lambda^{r,l} - \int_{(r',l')\geq(r,l)} \gamma(r',l')\Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'}, \\ \hat{v}_L(r,l) &:= v_L(r,l) + \int_{(r',l')\leq(r,l)} \Lambda_{\mathrm{d}r',\mathrm{d}l'}^{r,l} - \int_{(r',l')\geq(r,l)} \Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} &(20)\\ &= v_L(r,l) + \Lambda^{r,l} - \Lambda_{r,l},\end{aligned}$$

where we let $\Lambda^{r,l} := \int_{(r',l')\leq(r,l)} \Lambda_{\mathrm{d}r',\mathrm{d}l'}^{r,l}$ be the marginal measure of $\Lambda_{r,l}^{r',l'}$ on outgoing ICs from type (r,l) to any lower type, and let $\Lambda_{r,l} := \int_{(r',l')\geq(r,l)} \Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'}$ be the marginal measure of $\Lambda_{r,l}^{r',l'}$ on incoming ICs from any higher type to (r,l).

We now write down the sufficient conditions for $\Lambda_{r,l}^{r',l'}$ to certify the optimality of (3). Partition E_2

into the following two sets depending on the allocation assigned by the proposed mechanism (3),

$$E_{00} := \{(r, l) \in E_2 : r < \underline{e}\},$$
$$E_{10} := \{(r, l) \in E_2 : r \geq \underline{e}\}.$$

Types in E_{10} are given allocation $(a_R, a_L) = (1, 0)$ and types in E_{00} are given allocation $(a_R, a_L) = (0, 0)$ by the mechanism in (3). Therefore, the goal is to find some $\Lambda_{r,l}^{r',l'}$ so that

$$\begin{aligned} &\hat{v}_R(r, l) \geq 0, \hat{v}_L(r, l) \leq 0, \forall (r, l) \in E_{10}, \\ &\hat{v}_R(r, l) \leq 0, \hat{v}_L(r, l) \leq 0, \forall (r, l) \in E_{00}, \end{aligned} \tag{21}$$

It remains to show that such a $\Lambda_{r,l}^{r',l'}$ exists.

Existence of a Multiplier as a Transport Problem The existence problem of $\Lambda_{r,l}^{r',l'}$ is a transport problem. To see this, according to (19) and (20), any mass that $\Lambda_{r,l}^{r',l'}$ places on an IC constraint from (r', l') to (r, l) decreases the value of $v_R(r, l)$ (respectively, $v_L(r, l)$) and increases the value of $v_R(r', l')$ (respectively, $v_L(r', l')$) by the same amount.

Our problem then becomes, given functions v_L and v_R, whether there exists a transport map $\Lambda_{r,l}^{r',l'}$ that satisfies (17) and (21), and never places mass on (r', l', r, l) such that $(r', l') \geq (r, l)$ such that $(r, l) \in E_{00}$ and $(r', l') \in E_{10}$ because ICs can only bind within E_{10} and E_{00}.

It suffices to construct separate transport plans within E_{10} and E_{00} to satisfy (17) and (21). This is because $\Lambda_{r,l}^{r',l'}$ is not allowed to transport values across E_{10} and E_{00}: The IC constraints only bind within each set for the mechanism in (3).[15]

Types in E_{10} We partition E_{10} based on the sign of $v_R(r, l)$. For any $(r, l) \in E_{10}$, $v_L(r, l) \leq 0$ due to Lemma 3, and $v_R(r, l)$ can either be positive or negative. Define $E_{10}^- := \{(r, l) \in E_{10} : v_R(r, l) < 0\}$. This is the set of types with a negative v_R. Let $r_1 := \sup\{r : \exists l, (r, l) \in E_{10}^-\}$. r_1 is the maximum level of r evidence for any type in E_{10}^-. Let $E_{10}^+ := \{(r, l) \in E_{10} : v_R(r, l) \geq 0, r \leq r_1\}$. Due to assumption 3.1, the iso-value line $u(r, l) = 0$ is always upwards sloping. E_{10}^+ is always to the left of E_{10}^-.

We want to transport the positive values of $v_R(r, l)$ from types $(r, l) \in E_{10}^+$ to the negative values of $v_R(r', l')$ from types $(r', l') \in E_{10}^-$ to satisfy (21). We will show that there exists a transport map that satisfies (21) and only moves values from $(r', l') \in E_{10}^-$ to any type $(r, l) \in E_{10}^+$ with $(r, l) \leq (r', l')$. This means only the IC constraints from $(r', l') \in E_{10}^-$ to $(r, l) \in E_{10}^+$ with $(r, l) \leq (r', l')$ can bind. All other constraints are slack.

[15] For any $(r', l') \geq (r, l)$ with $(r', l') \in E_2^{10}$ and $(r, l) \in E_2^{00}$, the IC from (r', l') to (r, l) never binds because type (r', l') gets a strictly smaller utility by mimicking type (r, l).

For (21) to hold, we need for types $(r', l') \in E_{10}^{-}$,

$$\hat{v}_R(r', l') = v_R(r', l') + \gamma(r', l')\Lambda^{r',l'} \geq 0, \tag{22}$$

$$\hat{v}_L(r', l') = v_L(r', l') + \Lambda^{r',l'} \leq 0, \tag{23}$$

and for types $(r, l) \in E_{10}^{+}$,

$$\hat{v}_R(r, l) = v_R(r, l) - \int\limits_{(r',l')\geq(r,l)} \gamma(r', l')\Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} \geq 0, \tag{24}$$

$$\hat{v}_L(r, l) = v_L(r, l) - \Lambda_{r,l} \leq 0. \tag{25}$$

We now show that there exists a transport plan $\Lambda_{r,l}^{r',l'}$ such that (22)-(25) hold and it only involves moving v_R values from $(r, l) \in E_{10}^{+}$ to $(r', l') \in E_{10}^{-}$ with $(r', l') \geq (r, l)$. We want some $\Lambda_{r,l}^{r',l'}$ such that (22) binds for any $(r', l') \in E_{10}^{-}$, that is, move just enough positive values of v_R from E_{10}^{+} to E_{10}^{-} so that $\hat{v}_R(r', l') = 0$ for any $(r', l') \in E_{10}^{-}$. If such $\Lambda_{r,l}^{r',l'}$ exists, (23) is then automatically satisfied because $v_R(r', l') - \gamma(r', l')v_L(r', l') \geq 0$. Moreover, (25) holds due to Lemma 3. As for (24), it is implied by the following inequality and (25)[16]

$$\int\limits_{(r',l')\geq(r,l)} \left[\gamma(r', l') - \gamma(r, l)\right] \Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} \leq v_R(r, l) - \gamma(r, l)v_L(r, l). \tag{26}$$

(26) always holds by construction. To see this, the right hand side of (26) is negative due to Lemma 4. The left hand side of (26) is always positive due to Lemma 2, which implies that $\gamma(r', l') - \gamma(r, l) < 0$ for any $(r', l') \in E_{10}^{-}$ and $(r, l) \in E_{10}^{+}$ with $(r, l) \leq (r', l')$.

It remains to show that there exists a $\Lambda_{r,l}^{r',l'}$ that shifts just enough values of v_R from E_{10}^{+} to E_{10}^{-} so that (22) binds. By Strassen's theorem (1965), such transport plan $\Lambda_{r,l}^{r',l'}$ exists if $v_R^{+}(r, l) := \max\{v_R(r, l), 0\}$ is first order stochastically dominated by $v_R^{-}(r, l) := \max\{-v_R(r, l), 0\}$. To be precise, view v_R^{+} and v_R^{-} as the densities of measures μ_R^{+} and μ_R^{-} on E_{10}. We need that μ_R^{+} to be first order stochastically dominated by μ_R^{-}. This ensures that there is enough positive mass in μ_R^{+} to be shifted up to μ_R^{-}. In fact, we have to make sure μ_R^{+} has enough total mass. Indeed, we always have

$$\int_{E_{10}} v_R^{+}(r, l)\mathrm{d}r\mathrm{d}l \geq \int_{E_{10}} v_R^{-}(r, l)\mathrm{d}r\mathrm{d}l$$

since otherwise the principal would have optimally failed all types in state R. Observe also that the above inequality may be strict, in which case we have excess positive mass. This never becomes an issue when we apply Strassen's theorem because we can always leave the excessive positive mass untouched.

To check for first order stochastic dominance in this two-dimensional space, typically we have to

[16] To see this, take (26) plus $\gamma(r, l)$ times (25) and we have (24).

show that the integral of v_R^+ over any lower set S is larger than that of v_R^- (see Theorem 4). In this problem, however, it suffices to check lower sets of the form $S^{r',l'} := (r,l) \in E_2 : (r,l) \leq (r',l')$. This comes from our assumption 3, which implies that the iso-value line $v_R(r,l) = 0$ is always upward sloping with $v_R < 0$ to its left and $v_R > 0$ to its right.

Therefore, it suffices to verify that, for any $(r',l') \in E_{10}^-$, there is more positive values to the southwest of (r',l') in E_{10}^+ than the negative values to the southwest of (r',l') in E_{10}^-. Formally, we need

$$\int_{\underline{e}}^{r'} \int_0^{l'} v_R^+(r,l)\mathrm{d}l\mathrm{d}r \geq \int_{\underline{e}}^{r'} \int_0^{l'} v_R^-(r,l)\mathrm{d}l\mathrm{d}r.$$

Plug in v_R^+ and v_R^- and split the integral into E_{10}^+ and E_{10}^-, the above becomes

$$\int_{\underline{e}}^{r'} \int_0^{l'} v_R(r,l)\mathrm{d}l\mathrm{d}r \geq 0$$

Since $v_R(r',l') < 0$ for any $(r',l') \in E_{10}^-$, the above inequality is implied by

$$\int_{\underline{e}}^{r'} \int_0^{\bar{l}(r)} v_R(r,l)\mathrm{d}l\mathrm{d}r \geq 0 \tag{27}$$

which sets l' to be the upper bound $\bar{l}(r) := \max\{l : (r,l) \in E_2\}$ to include all negative v_R.

(27) must hold due to the optimality of $\underline{e}$. To see this, the integral in (27) is the principal's value from passing in state R every type $(r,l) \in E_2$ with $\underline{e} \leq r \leq r'$. This integral must be positive, otherwise the principal can fail in state R every type $(r,l) \in E_2$ with $\underline{e} \leq r \leq r'$ and obtain a strictly higher payoff, contradicting the optimality of $\underline{e}$.

Types in E_{00} The construction for E_{00} is similar. We partition E_{00} based on the sign of $v_R(r,l)$. For any $(r,l) \in E_{10}$, $v_L(r,l) \leq 0$ due to Lemma 3, and $v_R(r,l)$ can either be positive or negative. Define $E_{00}^+ := \{(r,l) \in E_{00} : v_R(r,l) > 0\}$. This is the set of types with a positive v_R. Let $r_2 := \inf\{r : \exists l, (r,l) \in E_{00}^+\}$. r_2 is the minimum level of r evidence for any type in E_{10}^+. Let $E_{00}^- := \{(r,l) \in E_{00} : v_R(r,l) \leq 0, r \geq r_2\}$. Due to assumption 3.1, the iso-value line $u(r,l) = 0$ is always upwards sloping. E_{00}^+ is always to the left of E_{00}^-.

We again want to transport the positive values of $v_R(r,l)$ from types $(r,l) \in E_{10}^+$ to the negative values of $v_R(r',l')$ from types $(r',l') \in E_{10}^-$ to satisfy (21). We will show that there exists a transport map that satisfies (21) and only moves values from $(r',l') \in E_{10}^-$ to any type $(r,l) \in E_{10}^+$ with $(r,l) \leq (r',l')$. This means only the IC constraints from $(r',l') \in E_{10}^-$ to $(r,l) \in E_{10}^+$ with $(r,l) \leq (r',l')$ can bind. All other constraints are slack.

For (21) to hold, we need for types $(r', l') \in E_{00}^-$,

$$\hat{v}_R(r', l') = v_R(r', l') + \gamma(r', l')\Lambda^{r', l'} \leq 0, \tag{28}$$

$$\hat{v}_L(r', l') = v_L(r', l') + \Lambda^{r', l'} \leq 0, \tag{29}$$

and for types $(r, l) \in E_{00}^+$,

$$\hat{v}_R(r, l) = v_R(r, l) - \int\limits_{(r', l') \geq (r, l)} \gamma(r', l')\Lambda_{r,l}^{\mathrm{d}r', \mathrm{d}l'} \leq 0, \tag{30}$$

$$\hat{v}_L(r, l) = v_L(r, l) - \Lambda_{r,l} \leq 0. \tag{31}$$

We now show that there exists a transport plan $\Lambda_{r,l}^{r',l'}$ such that (28)-(31) hold and it only involves moving v_R values from $(r, l) \in E_{00}^+$ to $(r', l') \in E_{00}^-$ with $(r', l') \geq (r, l)$. We want some $\Lambda_{r,l}^{r',l'}$ such that (30) binds for any $(r, l) \in E_{10}^+$, that is, move just enough positive values of v_R from E_{00}^+ to E_{00}^- so that $\hat{v}_R(r, l) = 0$ for any $(r, l) \in E_{00}^+$. (31) holds by construction. We must also show that such $\Lambda_{r,l}^{r',l'}$ respects (28) and (29). It suffices to consider (28) because (29) is implied by (28) due to Lemma 4.

Again by Strassen's theorem (1965), a transport plan $\Lambda_{r,l}^{r',l'}$ that makes (30) binding and satisfies (28) exists if a similar first order stochastic condition holds. A similar argument as the E_{10} case implies that, a sufficient condition for FOSD is, for any $(r, l) \in E_{00}^+$, there is more negative mass to the northeast of (r, l) in E_{00}^- than the positive mass to the northeast of (r, l) in E_{00}^+. Formally, we need

$$\int_r^{\underline{e}} \int_l^{\bar{l}(r)} v_R^+(r', l')\mathrm{d}l'\mathrm{d}r' \geq \int_r^{\underline{e}} \int_l^{\bar{l}(r)} v_R^-(r', l')\mathrm{d}l'\mathrm{d}r'.$$

Plug in $v_R^+(r, l) := \max\{v_R(r, l), 0\}$ and $v_R^-(r, l) := \max\{-v_R(r, l), 0\}$, and split the integral into E_{00}^+ and E_{00}^-, the above becomes

$$\int_r^{\underline{e}} \int_l^{\bar{l}(r')} v_R(r', l')\mathrm{d}l'\mathrm{d}r' \leq 0$$

Since $v_R(r, l) > 0$ for any $(r, l) \in E_{00}^+$, the above inequality is implied by

$$\int_r^{\underline{e}} \int_0^{\bar{l}(r')} v_R(r', l')\mathrm{d}l'\mathrm{d}r' \leq 0, \tag{32}$$

which sets $l = 0$ to include all positive v_R.

(32) must hold due to the optimality of $\underline{e}$. Again, the integral in (32) is the principal's value from passing in state R every type $(r', l') \in E_2$ with $r \leq r' \leq \underline{e}$. This integral must be negative, otherwise the principal can pass in state R every type $(r', l') \in E_2$ with $r \leq r' \leq \underline{e}$ and obtain a strictly higher payoff, contradicting the optimality of $\underline{e}$.

B.4 Proofs for Section 5

Assumptions 6 imply Assumptions 3.

The proof consists entirely of algebra. The details are presented below.

Proof. *Notation.* For any function – $u, f_q|_{q\in\{G,B\}}, \overline{f}, p$ – when we use it without an argument, the argument is (r, l). When we superscript it with c, it denotes the same function, but with the argument (l, r). We use the usual notation that for any function $\phi : E \to \mathbb{R}$, $\phi_e = \frac{\partial\phi}{\partial e}$.

Let $\xi := \frac{1-g}{g}$. Recall that $\alpha = u_B\frac{1-g}{g}$.

Assumptions 6 $\Rightarrow$ Assumptions 3.1 For $e \in \{r, l\}$, algebra shows that:

$$\frac{\partial u}{\partial e} > (<)0 \iff (\alpha + \xi) f_G f_B \left(\frac{f_{Ge}}{f_G} - \frac{f_{Be}}{f_B}\right) > (< 0)$$

When $e = r$, the first part of Assumption 6.2 is sufficient for the above to hold. When $e = l$, the second part of Assumption 6.2, combined with Assumption 6.1, is sufficient for the above.

Assumptions 6 $\Rightarrow$ Assumptions 3.2. Essentially using equation (12), all we have to show is isobelief curves are steeper than iso-utility (of the principal) curves, in the $l - r$ plane:

$$\left.\frac{dr}{dl}\right|_{u=\text{constant}} \leq \left.\frac{dr}{dl}\right|_{p=\text{constant}}$$

Algebra shows,

$$\left.\frac{dr}{dl}\right|_{p=\text{constant}} = \frac{\gamma\overline{f}^c_l - \overline{f}_l}{\overline{f}_r - \gamma\overline{f}^c_r}$$

Also, using expressions for u_l and u_r derived in the previous step,

$$\left.\frac{dr}{dl}\right|_{u=\text{constant}} = -\frac{u_l}{u_r} = \frac{\left(\frac{f_{Bl}}{f_B} - \frac{f_{Gl}}{f_G}\right)}{\left(\frac{f_{Gr}}{f_G} - \frac{f_{Br}}{f_B}\right)}$$

Hence,

$$\begin{aligned}\left.\frac{dr}{dl}\right|_{p=\text{constant}} &\geq \left.\frac{dr}{dl}\right|_{u=\text{constant}} \\ \Leftrightarrow \frac{\gamma\overline{f}^c_l - \overline{f}_l}{\overline{f}_r - \gamma\overline{f}^c_r} &\geq \frac{\left(\frac{f_{Bl}}{f_B} - \frac{f_{Gl}}{f_G}\right)}{\left(\frac{f_{Gr}}{f_G} - \frac{f_{Br}}{f_B}\right)}\end{aligned} \tag{33}$$

Using the fact that $\overline{f} = gf_G + (1-g)f_B$ (and similarly, $\overline{f}^c$), algebra shows that Assumption 6.3 is sufficient for (33).

□

B.5 Extensions

We assume at the outset, that $\max_{\tilde{e}} \left(\int_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right) > 0$. The case where this is not satisfied is already dealt with in the proof of our main characterization.

Implications of Assumption 7. Let $e^* := \max_e \phi(e,e) \leq 1$, $\bar{r}(l) := \sup\{r : (r,l) \in E\}$. Let $\bar{e}$ be defined by (10). Let l_1 and l_2 be the point where the curve $\{(r,l) \in E : u(l,r) = 0\}$ crosses the 45 degree line and the resource constraint $\phi(r,l) = 1$, respectively.

By Assumption 7, $\left(\int_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right)^+$ is single peaked, i.e.,

$$\left(\int_{l=\tilde{e}}^{e^*} \int_{r=l}^{\bar{r}(l)} u(l,r) dF(l,r) \right)^+ \quad \text{is single-peaked.}$$

Taking partial derivative w.r.t. l, we have the following equivalent version of Assumption 7.

$$\int_l^{\bar{r}(l)} u(l,r') f(l,r') dr' \leq 0, \forall l_1 \leq l \leq \bar{e}, \tag{34}$$

$$\int_l^{\bar{r}(l)} u(l,r') f(l,r') dr' \geq 0, \forall \bar{e} \leq l \leq l_2, \tag{35}$$

Here we use the fact that $\left(\int_{\min\{r,l\} \geq \tilde{e}, r \geq l} u(l,r) dF(l,r) \right)^+$ is maximized at $\tilde{e} = \bar{e}$, by the optimality of $\bar{e}$ within the class of mechanisms described in Theorem 2.

Note also that by Assumption 3.1, $u(r,l) > u(l,r)$ for all $r \geq l$. This gives us the following:

$$\int_l^{\bar{r}(l)} u(r',l) - u(l,r') f(r',l) dr' \geq 0, \forall l_1 \leq l \leq l_2 \tag{36}$$

Conditions (34)-(36) are what we use for our proof below.

Proof. The proof follows the same idea as the proof of Theorem 1 through multiplier construction. Again, it suffices to certify the optimality of the mechanism proposed in (3) in the relaxed problem

(16). We incorporate the IC constraints with the multiplier $\Lambda_{r,l}^{r',l'}$ which is a positive measure on $E_2 \times E_2 := \{(r',l',r,l) : (r',l'),(r,l) \in E_2\}$ that satisfies (17).

We now split E_2 into three parts based on the allocation (a_R, a_L) assigned by the proposed optimal mechanism. Define

$$\begin{aligned} E_{00} &:= \{(r,l) \in E_2 : r < \underline{e}\}, \\ E_{10} &:= \{(r,l) \in E_2 : r \geq \underline{e}, l < \bar{e}\}, \\ E_{11} &:= \{(r,l) \in E_2 : l \geq \bar{e}\}. \end{aligned}$$

Types in E_{00} are given allocation $(a_R, a_L) = (0,0)$, types in E_{10} are given allocation $(a_R, a_L) = (1,0)$, and types in E_{11} are given allocation $(a_R, a_L) = (1,1)$ by the mechanism in (9).

To certify the optimality of (9), we want to construct a transport plan $\Lambda_{r,l}^{r',l'}$ so that

$$\begin{aligned} &\hat{v}_R(r,l) \leq 0, \hat{v}_L(r,l) \leq 0, \forall (r,l) \in E_{00}, \\ &\hat{v}_R(r,l) \geq 0, \hat{v}_L(r,l) \leq 0, \forall (r,l) \in E_{10}, \\ &\hat{v}_R(r,l) \geq 0, \hat{v}_L(r,l) \geq 0, \forall (r,l) \in E_{11}. \end{aligned} \tag{37}$$

We can separately construct $\Lambda_{r,l}^{r',l'}$ within E_{00}, E_{10} and E_{11} because again the IC constraints only bind within each set for the proposed mechanism (9).

Types in E_{00} The construction for E_{00} is exactly the same as before. The part for E_{00} in the proof of Theorem 1 still applies. Introducing E_{11} does not change anything for E_{00}.

Types in E_{10} We partition E_{10} now based on the signs of both $v_R(r,l)$ and $v_L(r,l)$. Define

$$\begin{aligned} E_{10}^{--} &:= \{(r,l) \in E_{10} : v_R(r,l) < 0, v_L(r,l) \geq 0\}, \\ E_{10}^{++} &:= \{(r,l) \in E_{10} : v_R(r,l) \geq 0, v_L(r,l) < 0\}. \end{aligned}$$

Let $r_1 := \sup\{r : \exists l, (r,l) \in E_{10}^{--}\}$. r_1 is the maximum level of r evidence for any type in E_{10}^{--}. Similarly, let $l_1 := \inf\{l : \exists r, (r,l) \in E_{10}^{++}\}$. l_1 is the minimum level of l evidence for any type in E_{10}^{++}. Next, define

$$\begin{aligned} E_{10}^{r_1+-} &:= \{(r,l) \in E_{10} : v_R(r,l) \geq 0, v_L(r,l) \geq 0, r \leq r_1\}, \\ E_{10}^{l_1+-} &:= \{(r,l) \in E_{10} : v_R(r,l) \geq 0, v_L(r,l) \geq 0, l \geq l_1\}. \end{aligned}$$

$E_{10}^{r_1+-}$ and $E_{10}^{l_1+-}$ do not overlap. This is because (r_1, l_1) is the point where the iso-value curve $u(r,l) = 0$ crosses the 45 degree line. Also, there is no $(r,l) \in E_{10}$ with $v_R(r,l) < 0$ and $v_L(r,l) < 0$

to worry about due to symmetry.

We want to shift the positive values of v_R from $E_{10}^{r_1+-}$ to E_{10}^{--}, and the positive values of v_L from E_{10}^{++} to $E_{10}^{l_1+-}$ so that (37) holds. We can do this separately for the two pairs of sets.

To shift the positive values of v_R from $E_{10}^{r_1+-}$ to E_{10}^{--}, we can use the same construction as in the part proof of Theorem 1. Nothing has changed.

To shift the positive values of v_L from E_{10}^{++} to $E_{10}^{l_1+-}$ in a way that satisfies (37), we need types $(r', l') \in E_{10}^{l_1+-}$ to satisfy

$$\hat{v}_R(r', l') = v_R(r', l') + \gamma(r', l')\Lambda^{r',l'} \geq 0, \tag{38}$$

$$\hat{v}_L(r', l') = v_L(r', l') + \Lambda^{r',l'} \leq 0, \tag{39}$$

and types $(r, l) \in E_{10}^{++}$ to satisfy,

$$\hat{v}_R(r, l) = v_R(r, l) - \int\limits_{(r',l')\geq(r,l)} \gamma(r', l')\Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} \geq 0, \tag{40}$$

$$\hat{v}_L(r, l) = v_L(r, l) - \Lambda_{r,l} \leq 0. \tag{41}$$

This is where the additional assumption 7 kicks in. Assumption 7 implies the following,

$$\int_l^{\bar{r}(l)} v_L(l, r')dr' \leq 0, \forall l_1 \leq l \leq \bar{e}, \tag{42}$$

$$\int_l^{\bar{r}(l)} (v_R(r', l) - \gamma(r', l)v_L(l, r'))dr' \geq 0, \forall l_1 \leq l \leq \bar{e}. \tag{43}$$

(42) says that there is enough negative values of v_L to take in all the positive values of v_L in E_{10}^{++} along every l. This implies that the transport plan can be constructed within each l to satisfy (39) and (41) simultaneously. (38) holds because $v_R(r', l') \geq 0$ for $(r', l') \in E_{10}^{l_1+-}$. It remains to check that (40) holds.

(43) implies that when we use the within-l transport plan to make (41) bind, we can always make (40) hold. To see this, (43) says that there is enough positive values of v_R for types in E_{10}^{++} so that when they shift the negative values of v_L upwards, the virtual values $\hat{v}_R$ stays positive.

Types in E_{11} We partition E_{11} based on the sign of $v_L(r, l)$. Define $E_{11}^{-} := \{(r, l) \in E_{11} : v_L(r, l) < 0\}$. Let $l_2 := \sup\{l : \exists r, (r, l) \in E_{11}^{-}\}$. l_2 is the maximum level of l evidence for any type in E_{11}^{-}. Define $E_{11}^{+} := \{(r, l) \in E_{11} : v_L(r, l) \geq 0, l \leq l_2\}$. Note that for any $(r, l) \in E_{11}$, $v_R(r, l) \geq 0$.

We want to transport the positive values of $v_L(r, l)$ from types $(r, l) \in E_{11}^{+}$ to the negative values of $v_R(r', l')$ from types $(r', l') \in E_{11}^{-}$ to satisfy (37). We will show that there exists a transport map that satisfies (37) and only moves values from $(r',l') \in E_{11}^{-}$ to any type $(r,l) \in E_{11}^{+}$ with $(r,l) \leq (r',l')$. This means only the IC constraints from $(r',l') \in E_{11}^{-}$ to $(r,l) \in E_{11}^{+}$ with $(r,l) \leq (r',l')$ can bind. All other constraints are slack.

For (37) to hold, we need for types $(r',l') \in E_{11}^{-}$,

$$\hat{v}_R(r',l') = v_R(r',l') + \gamma(r',l')\Lambda^{r',l'} \geq 0, \tag{44}$$

$$\hat{v}_L(r',l') = v_L(r',l') + \Lambda^{r',l'} \geq 0, \tag{45}$$

and for types $(r,l) \in E_{11}^{+}$,

$$\hat{v}_R(r,l) = v_R(r,l) - \int\limits_{(r',l')\geq(r,l)} \gamma(r',l')\Lambda_{r,l}^{\mathrm{d}r',\mathrm{d}l'} \geq 0, \tag{46}$$

$$\hat{v}_L(r,l) = v_L(r,l) - \Lambda_{r,l} \geq 0. \tag{47}$$

This is again where we need the additional assumptions. Assumption 7 implies that

$$\int_l^{\bar{r}(l)} v_L(l,r')dr' \geq 0, \forall \bar{e} \leq l \leq l_2, \tag{48}$$

$$\int_l^{\bar{r}(l)} (v_R(r',l) - \gamma(r',l)v_L(l,r'))dr' \geq 0, \forall \bar{e} \leq l \leq l_2. \tag{49}$$

(48) says that there is enough positive values of v_L to fill in all the negative values of v_L in E_{11}^{-} along every l. This implies that the transport plan can be constructed within each l to satisfy (45) and (47) simultaneously. (44) holds again because $v_R(r',l') \geq 0$ for $(r',l') \in E_{11}^{-}$. It remains to check that (46) holds.

(49) implies that when we use the within-l transport plan to make (47) bind, we can always make (46) hold. To see this, (49) says that there is enough positive values of v_R for types in E_{11}^{+} so that when they shift the values of v_L upwards, the virtual values $\hat{v}_R$ stays positive.

This completes the proof. □