Screening Knowledge*

Frequently updated. Please click here for the latest version.

Sulagna Dasgupta

December 12, 2023

Abstract

A principal (she) tests an agent's (he) knowledge of a subject matter. She has preferences over his unobserved quality, which is correlated with his knowledge. Modeling knowledge as beliefs over an unknown state, I show that optimal tests are simple: They take the form of True-False, weighted True-False or True-False-Unsure, regardless of the principal's preferences, the distribution of the agent's beliefs, its correlation with his quality or his knowledge thereof. The need to elicit knowledge forces the principal to trade-off the efficacy of the test in terms of *whom* it rewards, against *how much* it rewards them. The optimal resolution of this trade-off may lead to a partial penalty for an "obvious" answer even if it is correct, a partial reward for a "counter-intuitive" answer even if it is incorrect, or a reward for admitting ignorance. When the principal can pick the subject matter, she picks one that admits no such obvious answers. In this case, the highly prevalent True-False test is always optimal, regardless of principal's preferences, agent's learning, or the specific optimal choice of the subject matter.

JEL Classification: D82
Keywords: Knowledge-based screening, scoring rules, test design

*I am most grateful to Ben Brooks, Daniel Rappaport, Doron Ravid and Phil Reny for their guidance and support. This project has also benefited tremendously from conversations with Emir Kamenica, Joe Root, Alex Frankel, Marina Halac, Andreas Kleiner, Nicholas Lambert, Idione Meneghel, Ilia Krasikov, Zizhe Xia, Lenka Fiala, Deepal Basak and various audiences at the University of Chicago. All errors are my own.

1 Introduction

> "Ignorance is preferable to error, and he is less remote from the truth who believes nothing than he who believes what is wrong."
>
> Thomas Jefferson

Traditionally, many standardized testing institutions across the world seem to have taken the above concept seriously – by penalizing wrong answers more than *ignorance*, i.e. questions left unanswered. For example, the Advance Placement and SAT I examinations employed "negative marking" – deducting a fraction of a point for each wrong answer while awarding a point for each correct answer and ignoring unanswered questions – until recently.[1] However, in recent times there is a move away from negative marking – the College Board removed penalties for wrong answers on Advanced Placement examinations in 2011, and on the SAT I tests in 2014. In 2015, testing authorities in Chile removed these penalties from the University Selection Test.[2] While the equity impacts of these changes are largely considered to be positive (Baldiga, 2014; Saygin and Atwater, 2021; Coffman and Klinowski, 2020), how do they affect the core objective of these tests, which is to measure the quality of "college readiness"?

More broadly, in the context of knowledge-based evaluation schemes – such as the standardized tests mentioned above, in-class exams or job interviews – why should or shouldn't a candidate be rewarded for admitting ignorance? Zooming out even more, how best to design such tests? Does the universally common structure of responders giving *an* answer to a question – instead of more open-ended formats we can imagine, like expressing their views on how likely they think various possible answers are – sacrifice efficacy? How to ensure that the tests are not "gamed"?[3]

To answer these questions, I develop a theory of screening knowledge, recognizing the fact that a test-taker's knowledge is his private information. I consider a

[1] Other large scale standardized tests which have traditionally employed negative marking include the University Selection Test (Prueba de Selección Universitaria) in Chile, the medical entrance test Konkoor in Iran, the Higher Education Examination Undergraduate Placement Examination in Turkey, exams in the Ghent University system in Belgium and the MBA entrance Common Admission Test, the engineering entrace IIT-JEE, as well as the accountancy exam Common Proficiency Test in India. (Coffman and Klinowski, 2020; Akyol et al., 2016; Lesage et al., 2013).

[2] See Coffman and Klinowski (2020).

[3] A common example of profitably gaming the test is the following. A commonly suggested strategy by test prep coaches in tests with negative marking, is to randomly guess, instead of leaving the question, if the test-taker is able to eliminate at least one answer. For an expected points maximizer who is indifferent among all but one options, this strategy strictly dominates the desired behavior of leaving the question, for the most common negative marking scheme known as "$\frac{1}{n-1}$", i.e. where there are n options to each question and $\frac{1}{n-1}$ points are deducted for each wrong answer (Karandikar, 2010).

model where a test-designer – such as a teacher or interviewer – tests a candidate's knowledge of a single binary fact. She can choose from all possible tests with one restriction – the test must incentivize the candidate to reveal his true knowledge. Depending on what the test is trying to measure, the designer may have different preferences over knowledge. In many natural settings, she wants to reward those who are knowledgable and penalize those who are not.[4] The key insight from this framework, however, is that incentive compatibility forces her to trade off between the sharpness of this reward scheme on the extensive margin – *whom* it rewards – and that on the intensive margin – *how much* it rewards them.

I show that resolving this trade-off in a way that is optimal for the designer's objective of discerning candidate quality, can lead to potentially surprising tests which reward ignorance. This provides a basis for the negative marking in standardized tests, as discussed above. These results inform ongoing policy debates about the benefits of negative marking in large scale standardized tests.

I now discuss the model and results in more detail.

Our model is as follows. There is a principal (test-designer, *she*) who wants to decide whether to pass or fail an agent (test-taker, *he*). He has some underlying *quality*, unobserved by the principal, which we equate to her payoff if he is passed. Quality can be positive or negative. There is a binary state, representing a factual subject matter. The principal can test the agent's *knowledge* – modeled as his belief about the state – to decide whether to pass him. For that, she commits to a *mechanism*, which we interchangeably refer to as a *test*. A mechanism maps the agent's reported belief and *each* realized state – which the principal can precisely verify ex-post[5] – to a passing probability. The agent's quality is correlated (not necessarily positively) with the precision of his information about the state. Hence, observing his beliefs in conjunction with the state lets the principal form posterior estimates of his quality, which dictate her preferences over beliefs. The agent wants to maximize his passing probability.[6] There are no monetary transfers.

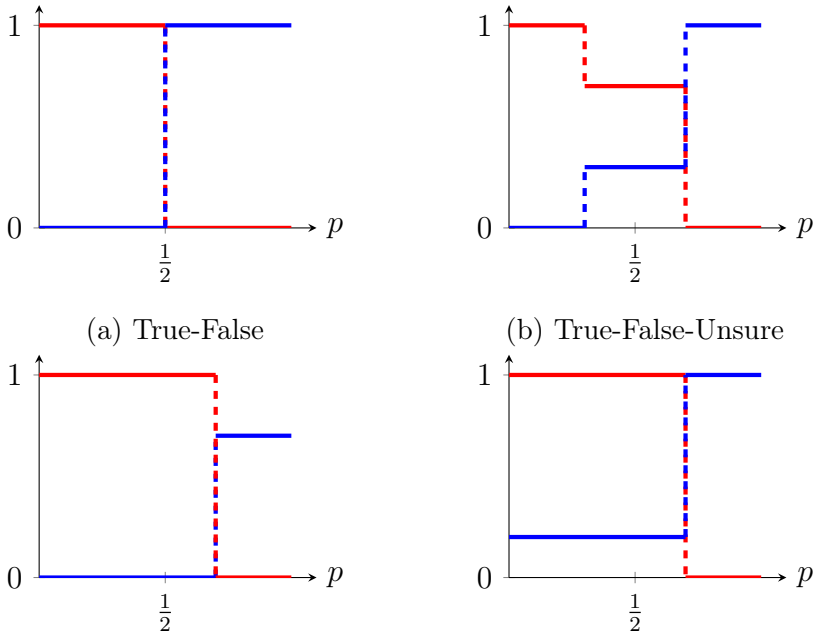
I begin by characterizing the class of optimal tests. Even though the principal has the flexibility to make the reward scheme as sensitive to the agent's knowledge as

[4]While intuitively we might assume more knowledge should always be preferred to less, our model requires no such monotonicity assumptions. Moreover, Section 2 provides a natural class of examples where even though the test-designer wants to reward "better" test-takers, who are also more informed, her induced preferences over knowledge may turn out to be non-monotonic.

[5]In Section 7.2.1 we consider the alternative timing where the principal observes the state *ex-ante* instead, i.e., before choosing the mechanism. Our main results remain qualitatively unaltered across a natural class of equilibria of the informed principal game Myerson (1983) which ensues in this case.

[6]The agent's probability of passing can be interpreted as number of points as well, if we assume the agent is an expected-points maximizer. For how our conclusions could change for other natural classes of the agent's preferences over points, see Section 7.5.

she likes, potentially rewarding each belief differentially according to its correctness, optimal tests fall within a simple class. In particular, they take the familiar form of a True-False, weighted True-False or True-False-Unsure test (See Table 1 for their precise definitions) regardless of the principal's preferences over quality, the distribution of the agent's beliefs, its correlation with his quality or his knowledge thereof (Theorem 1). The direct mechanisms depicted in Figure 1, provides an exhaustive enumeration of the possible types of optimal tests, in terms of their qualitative features.[7]



(a) True-False (b) True-False-Unsure

(c) weighted True-False (d) weighted True-False

Figure 1: Types of optimal tests

p denotes the agent's belief that the state is T. The blue and red curves denote passing probability in state T and F respectively.

[7]To be very precise, the class of optimal tests also includes *trivial* tests not included in Figure 1 – namely, under which the agent is passed or failed regardless of the state or his beliefs.

		Chosen Answer	
		T	F
Correct	T	x	y
Answer	F	0%	100%

(a) True-False, weighted True-False

		Chosen Answer		
		T	F	U
Correct	T	100%	0%	z_1
Answer	F	0%	100%	z_0

(b) True-False-Unsure

Table 1: Indirect implementations of optimal tests

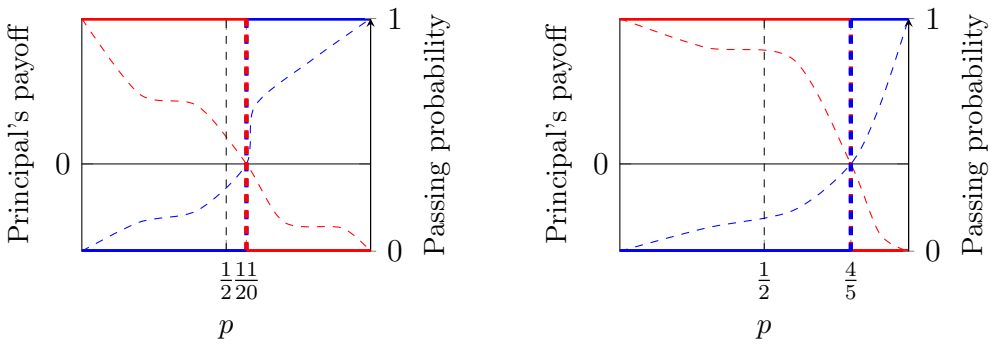
A test evaluates the agent's knowledge of whether a given statement is True (T) or False (F). The tables capture the following natural indirect implementation of optimal tests. The optimal test[8] gives the test-taker two options, mirroring the possible answers. It may, *in addition*, give him the option of declaring himself *Unsure (U)* (Table 1(b)). The cell entries capture percentages of the full credit earned for each combination of the correct answer and the chosen option.

Table 1(a) captures the generalized structures of True-False and weighted True-False tests. We call a test with $x = 100\%$ and $y = 0\%$ a *True-False* test. We show that if $x \in (0, 1)$ (respectively, $y \in (0, 1)$) under an optimal test, we must have $y = 0\%$ (respectively, $x = 100\%$). Tests with $x \in (0, 1)$ or $y \in (0, 1)$ are called *weighted True-False* tests. Finally, tests of the structure in Table 1(b) are called True-False-Unsure tests, $z_1, z_0 \in (0, 1)$.

The direct mechanisms leading to these tests are detailed in Figure 1.

Two features of the optimal tests warrant emphasis. First, agents giving the correct answer are sometimes failed – even the ones with the most precise possible correct beliefs – and vice versa (Figures 1c and 1d). Second, agents can be rewarded for admitting ignorance, i.e. choosing the "Unsure" option in a True-False-Unsure test (Figure 1b). If a regularity condition is satisfied the second feature does not arise (Theorem 2.A). This condition is mathematically similar to Myersonian regularity – it is equivalent to the increasingness of a *virtual value* function that arises in our setting (Proposition 1). However, unlike Myersonian regularity, its violation is arguably rather natural in a broad class of settings (Proposition 3). I postpone the discussion of such settings until a few paragraphs later, focusing on regular settings for now, to highlight the fundamental incentive issue in the knowledge screening problem, and to build intuition for the first of the aforementioned features.

[8]To be precise, first, we do not claim that any optimal test within the aforementioned class is uniquely optimal, and secondly, Tables 1 do not include the trivial tests included in the class of optimal tests, which pass or fail the agent without screening.



(a) Prior close to a half

(b) A skewed prior

Figure 2: Examples of the principal's preferences and corresponding first-best tests

The figures feature double axes. The agent's belief (p) that the state is T is along the horizontal axis. The dotted blue and red curves plot the posterior expected quality in states T and F respectively, on the left vertical axis. Superimposed on this plot, the solid blue and red curves plot the passing probability in states T and F respectively, under the first-best test, on the right vertical axis.[9]

To understand the role of incentives, we would start from the *first-best* – the benchmark where the principal can observe the agent's beliefs.[10] As examples, we consider the preferences of the principal depicted in Figures 2a and 2b. In these examples, the posterior quality is positive (respectively, negative) for beliefs which are more biased towards the correct (respectively, incorrect) state than the prior, and is zero at the prior – which is 55% and 80 % that the state is T, in figures 2a and 2b respectively. Under the first-best, she passes a belief if and only if its posterior quality is positive.

We first consider Figure 2a, where the prior is "moderate" – 55%. The first-best test, as depicted in Figure 2a is not incentive compatible, for the following reason. Essentially, under the first-best test, agents are required to "guess the state" and are passed if and only if they guess correctly, but they "should" guess T only if they are at least 55 % sure. Clearly, this test induces agents to guess the state they think is more likely. Hence, those with beliefs in between one half and 55 %, deviate to guessing T instead of F.

The only incentive compatible test which preserves the desired first-best feature

[9]Note that Figures 2 provide *schematic* representations of possible principal preferences. In particular, her interim values (interim quality) in each state reaching zero at the same belief, which is also the prior, is a *non-generic* feature of such preferences. I still use these features in these examples, in order to distill the most important forces of incentives, and how they vary with the prior – my main focus in this paper.

[10]Note that under our first-best, the principal observes only the agent's belief, not any other information such as his quality, including in the version of our model where the agent has other private information. See Section 7.1 for more details.

of being bang-bang,[11] is the one with a belief threshold of one half. It turns out, for a moderate prior, as in Figure 2a, that is the best the principal can do – simply shifting the belief threshold to one half (Theorem 2.A). This gives rise to the optimal test in Figure 1a – the ubiquitous True-False test. In this particular case, this happens because the principal's ideal threshold is "close enough" to one half. By this reasoning, a threshold of one half must be optimal for a band of ideal thresholds of the principal, i.e. the True-False test must arise "generically". This is one way we provide a basis for its prevalence.[12]

Now we consider a more extreme prior – 80 % for T – as in Figure 2b. For restoring incentives, shifting the corresponding highly unbalanced threshold all the way to one half is costlier to the principal in this case, than keeping the threshold unbalanced, but adjusting the passing probabilities in a way which prevents deviations. Such adjustments must result in smaller overall rewards for "guessing" T than for guessing F, so that only those sufficiently biased towards state T, guess T, thereby implementing the desired skewed threshold. Optimally, this can be achieved by either bringing up the passing rate for wrongly guessing F (Figure 1d) or bringing down that for correctly guessing T (Figure 1c), so as to make the threshold belief indifferent. Which of these two adjustments is optimal is determined by whether the principal would pass or fail "by default", if she had to take that decision without screening, i.e., whether the prior expected quality is positive or negative (Proposition 2).

Thus, the key trade-off the test-designer faces is one between the efficacy of the reward scheme on the intensive and extensive margins. As outlined above, the distortions of potential penalties for the correct answer or rewards for the wrong one arise because she trades off choosing her preferred belief threshold – which determines *which* beliefs are passed in each state (extensive margin) – against *how much* probability they are passed with, in each state, on its two sides (intensive margin).

As we saw above, for either type of adjustment, the a priori unlikely – or *counterintuitive* – answer is rewarded more vis a vis the a priori likely – or *obvious* – answer (Corollary 1). Moreover, degree of this premium increases as the obvious answer becomes more obvious, i.e., the prior grows more extreme (Theorem 2.B). This reflects the common feature of real world evaluation schemes which sometimes attach greater penalty to getting "obvious" questions wrong than to getting "trick" questions wrong.

We now turn to the case where the optimal test features a third option – *Unsure* (Figure 1b). Intuitively, a True-False-Unsure test is optimal in the following class of

[11] i.e., switches abruptly from 0 to 1.

[12] An additional basis for its prevalence arises by endogenizing the topic – modeled as endogenizing the prior – as described a few paragraphs later. See Section 6 for details.

settings. Suppose there are three types of signals – correct, wrong and uninformative. Better agents are much more likely to receive a correct signal and much less likely to receive a wrong one than worse agents, but all agents are almost similarly likely to receive an uninformative signal. Hence, while extreme beliefs act as strong and opposite (positive and negative) signals of quality to the principal in the two states, moderate beliefs provide only a weak signal of quality in either state. This induces the principal to keep the variation in rewards across states low for *unsure* agents – those with moderate beliefs – while varying them starkly for those with more extreme beliefs, as depicted in Figure 1b (formalized in Proposition 3).

Finally, we ask how the optimal test would look in the rather natural scenario where the principal can also choose the topic under testing. We model this as her choosing the prior. We show that the principal prefers a moderate prior – embodying greater a priori uncertainty – over an extreme one, though the optimal prior need not equal one half (Theorem 4). The broad intuition is that extreme priors "waste" information: The informativeness of a correct answer as a signal of quality in the a priori likely state falls too low because too much information is given away by the prior. Building on this insight, we further show that when the principal can choose the prior, the True-False test (Figure 1a) is optimal for all signal structures and preferences of the principal. This provides another basis for its prevalence in the real world.

Finally, we consider the case when the principal is endowed with the correct answer before the game begins. We show that our results remain qualitatively valid in any *undominated* equilibrium Myerson (1983) of the informed principal game which ensues in this case. (Proposition 7 and Theorem 5). Our main results also qualitatively hold in the natural case when the test-designer wants to maximize social welfare, instead of just screening efficacy (Proposition 9). We conclude by discussing the extent to which our results apply to multi-question tests (Proposition 10).

To summarize, this paper highlights the role of agency issues in designing knowledge-based tests for agents, when their knowledge generates value for the principal. The optimal tests are shown to take simple true-false/true-false-uncertain forms. The agency issues are shown to give rise to rewarding of not only correctness – as would happen under full information – but also *surprisingness* of answers. Moreover, we provide a basis for the ubiquitous simple True-False test by showing that it arises universally, whenever the test-designer can choose both the evaluation scheme and the question.

Related Literature. This paper relates to several different strands of literature. I share the core of my model with that on evaluation of strategic forecasters, most importantly Deb et al. (2018) and to a lesser extent, Chambers and Lambert (2021). In Deb et al. (2018), the authors develop a dynamic model for screening an election forecaster who privately observe signals about the election outcome which grow more precise with time. My model, though static, allows the principal to choose from the entire universe of tests, unlike theirs, where she is restricted to deterministic tests. The literature has also studied forecasters being evaluated by a passive market in settings without commitment (e.g. Ottaviani and Sørensen (2006)). Also see Marinovic et al. (2013) for a survey.

In their independent and concurrent work Deb et al. (2023) consider a joint screening-and-persuasion problem and find a similar characterization of the class of optimal mechanisms. However, while Deb et al. (2023) focus on the role of commitment, with their main result establishing the sufficiency of partial commitment in their setting, this paper focuses on the trade-off between effectiveness of screening on the intensive and extensive margins created by agency issues, and on the joint design problem of the prior and the mechanism.

The role of belief-based screening in my model relates it to the literature on *proper scoring rules*. The latter term describes mechanisms that incentivize an agent to reveal his true beliefs about an uncertain state. Much of the classical literature on proper scoring rules focuses on characterizing the set of incentive compatible scoring rules in general environments (McCarthy (1956), Osband and Reichelstein (1985), Lambert (2011), Abernethy and Frongillo (2012), while remaining agnostic about the designer's objectives. A notable exception is Li et al. (2022), most related to my work within this literature, who investigate how to optimize such rules, where the objective is to incentivize effort by the agent to acquire more precise signals.

At a broader level, this paper also relates to the literature on mechanism design with verification but without transfers. Like in my 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 my 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 et al. (2014) which features a similar multi-agent model, but with exact verification at a cost.

Thematically, though not as closely in terms of model or methods, my work is also related to the literature on how a receiver of information (in my 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 my paper, some of this literature finds *coarse* tests arising at the optimum (Rosar (2017), Harbaugh and Rasmusen (2018)).

2 A general model

There is a principal (she) and an agent (he). There is a binary state, $\omega \in \Omega := \{0, 1\}$, with prior probability $\pi \in (0, 1)$ of being equal to 1. The agent observes a signal about the state and forms beliefs. Let us denote the agent's belief that the state is 1 by $p \in P \subseteq [0, 1]$. We assume P is a compact interval, in particular, $P = [\underline{p}, \overline{p}]$. Wherever appropriate, we refer to the agent's belief as his *type*, capturing the fact that it is his private information and he is screened on it. Let $F(\cdot|\omega)$ denote the cumulative distribution of beliefs in state ω. The principal can take one of two actions: pass or fail the agent. The agent prefers to be passed. There are no transfers. Let $v_\omega(p)$ denote the principal's payoff from passing an agent with belief p in state ω. Her payoff from failing him is zero. We assume passing the agent hurts the principal with positive probability – otherwise the problem is trivial. Formally, we assume:

$$\max_{\omega} \int_{\{p: v_\omega(p) < 0\}} dF(p|\omega) > 0 \qquad \text{(Non-triviality)}$$

The above setup can arise from, for example, the principal (e.g. an employer) benefiting from the precision of the agent's (e.g. a job candidate) knowledge, i.e. v_ω increasing in p for $\omega = 1$ and decreasing for $\omega = 0$. However, we require no such assumptions on $\{v_\omega\}_{\omega \in \{0,1\}}$. At this stage we deliberately abstract away from the specifics of the principal's preference over the agent's beliefs as well as the latter's signal structure, because our characterization of the class of optimal mechanisms depends on none of these particulars. See Section 3 for an example and section 5.1 for a more general microfoundation of the above model.

The principal chooses a *mechanism* to determine passing decisions for agents. A mechanism is a tuple $\mathcal{M} := (M, a_1, a_0)$ where M is a set (of *messages*, to be sent by the agent) and $a_\omega : M \to [0, 1]$ is the passing probability or *passing rate* of the agent upon sending message $m \in M$, when the true state is ω. The principal is risk-neutral. She chooses this mechanism with the goal of maximizing her own ex-ante payoff.

The timing of the game is as follows:

1. The principal commits to a mechanism $\mathcal{M} = (M, a_1, a_0)$.

1. The agent observes his signal.

2. The agent makes his report $m \in M$ to the mechanism.

3. The state ω is observed by the principal.

4. Allocations are made according to $\mathcal{M}$ and payoffs are realized.

The same numbering of the first two stages of the game indicates that their ordering does not matter.

3 An illustrative example

A teacher evaluates a student based on his answer to whether a given factual statement is true (T) or false (F), using a pass-fail test. Hence, our unknown binary state is, whether the statement is actually True or False. We encode True as 1 and False as 0. Hence, going forward, our state is $\omega \in \{0, 1\}$. A student's preparation can be *High* (H) or *Low* (L) with equal probability, $\nu := \frac{1}{2}$. The teacher wants to pass only students with high preparation. In particular, let us assume she gets a payoff of $u_H := 2$ from passing the High type and $u_L = -1$ from passing the Low type.

We describe the student's learning process for any fact using the following signal structure with the unit interval, $T = [0, 1]$, as his signal space. The signal density of learning type $e \in \{H, L\}$ in state ω is $f_{e\omega} : T \to \mathbb{R}_+$. Let $f_{H1}(t) = 2t, f_{H0}(t) = 2(1-t), f_{L1}(t) = f_{L0}(t) = 1$.

In order to avoid giving away hints to the answer through her choice of question,[13] she promises to choose the question randomly from a question bank which contains a mix of true and false statements. Suppose a proportion π of these statements are actually true and this is common knowledge. Hence, π is the commonly held *prior belief* that the correct answer is *True*. Over the course of this example we will use two different values of π to highlight different aspects of the problem. First, let us assume $\pi = \frac{1}{2}$. We assume preparation and the correct answer are independent.

We also assume the student does not know whether his level of preparation is High or Low.[14] Hence he interprets his observed signal, $t \in [0, 1]$, using the *uniform average* of the two signal structures – the High and the Low type's – via Bayes rule. Let $\mu : T \to [0, 1]$ map each of his signals to his belief that the state is $\omega = 1$, i.e. that the given statement is actually true. By Bayes' rule we have:

[13]For example, consider the question: Is New York City the capital of New York state? (A)Yes (B)No. A completely uninformed person might get this question correct simply by reasoning that if the correct answer were the apparently "obvious" answer ((A)Yes, in this case) the teacher would not set such an easy question. Formally, this is an informed principal problem. See section 7.2.1 for details.

[14]While this is a simplifying assumption for this example, none of our main results depends on whether the agent knows his own ability.

$$\mu(t) = \frac{\pi(2t+1)}{\pi(2t+1) + (1-\pi)(2(1-t)+1)} = \frac{(2t+1)}{4} \qquad \text{(Belief)}$$

Using the above, the range of beliefs that the state is 1 is, $P = \left[\frac{1}{4}, \frac{3}{4}\right]$. Let $f(p|\omega)$ denote the density of the belief p in state $\omega \in \{0, 1\}$.

The teacher wants to construct a test to maximize her expected payoff. Even though the correct answer can only be True or False, she can structure the test in many ways — offering potentially uncountably many options (since the student can form uncountably many different beliefs) and an accompanying probability of passing when each of the options is chosen, for each correct answer.

Ideally, the teacher wants to offer high passing probability to a belief, if and only if it gives her a high payoff, if passed. What is her payoff from passing a given belief $p \in P$? That depends on the correct answer. Below we calculate it for the case when the true state is 1, i.e. the statement is actually true. Using Bayes' rule (we use $Pr(\cdot)$ to denote probability),

$$\begin{aligned} V_1(p) &= u_H Pr(H|p,1) + u_L Pr(L|p,0) \\ &= \frac{u_H Pr(p|H,1)Pr(H) + u_L Pr(L|p,0)pr(L)}{Pr(p|1)} \\ &= \frac{u_H \times \frac{1}{2} \times f_{1H}(\mu^{-1}(p))dt + u_L \times \frac{1}{2} \times f_{1L}(\mu^{-1}(p))dt}{f(p|1)dp} \end{aligned} \tag{1}$$

By (Belief) we have, if $p = \mu(t)$,

$$dp = \mu'(t)dt \implies dt = \frac{dp}{\mu'(\mu^{-1}(p))} = 2dp \tag{2}$$

The above tells us that the "weighted" value of belief p in state 1:

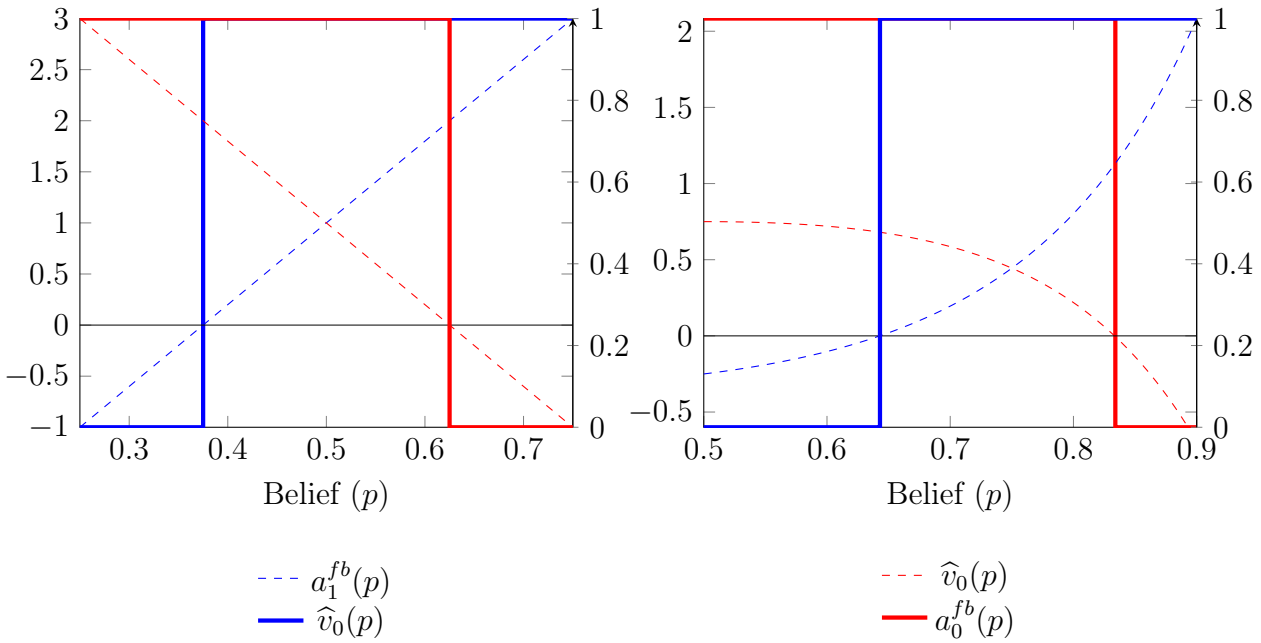
$$V_1(p)f(p|1)dp = \left(4p - \frac{3}{2}\right)dt = (8p-3)dp$$

The value function for state 0 can be similarly computed, and is the flipped version of $V_1(p)f(p|1)$, by symmetry.

Hence the "weighted" value functions are given by,

$$\begin{aligned} V_1(p)f(p|1) &= 8p - 3, \\ V_0(p)f(p|0) &= 8(1-p) - 3 = 5 - 8p. \end{aligned}$$

Let us denote the weighted value function in state ω by $\widehat{v}_\omega : P \to \mathbb{R}$, $\omega \in \{0, 1\}$, i.e. $\widehat{v}_\omega(p) = v_\omega(p)f(p|\omega)$ for all $p \in P$. These weighted value functions are are depicted by dashed lines in Figure 3a.



(a) Prior $\pi = \frac{1}{2}$ (b) Prior $\pi = \frac{3}{4}$

Figure 3: The teacher's value functions for symmetric and asymmetric priors

We first describe the teacher's optimal test if she could observe the students' beliefs but not their preparation levels – her "first best" – and then add back the incentive constraints to construct the optimal mechanism, as this is instructive in highlighting the role of incentives in this setting. This "first best" test is clearly given by $a_\omega^{fb}(p) = 1(v_\omega(t) \geq 0), \omega \in \{0, 1\}$, where $a_\omega^{fb}(p)$ denotes the passing rate of belief p under this test, when the true state is ω. This is depicted with solid lines in Figure 3a.

However, as we can see from Figure 3a, the first best is not implementable because the intermediate beliefs, $p \in \left[\frac{3}{8}, \frac{5}{8}\right]$, are passed for sure under this test, and therefore the extreme beliefs – those beyond this range – would deviate to this range, if this test is offered. Owing to the symmetry of the setting under the balanced prior, the familiar "Simple True-False" test, where there are two options mirroring the actual answers, and a student is passed if and only if his chosen answer is correct, is optimal in this case. The direct mechanism which represents this test is, of course, the one with a belief threshold of one half. Denoting the teacher's maximized payoff for prior π by $V(\pi)$:

$$
\begin{aligned}
V\left(\frac{1}{2}\right) &= (1-\pi)\int_{\frac{1}{4}}^{\frac{1}{2}} V_0(p)f(p|0)dp + \pi\int_{\frac{1}{2}}^{\frac{3}{4}} V_1(p)f(p|1)dp \\
&= 2\times\frac{1}{2}\int_{\frac{1}{2}}^{\frac{3}{4}} V_1(p)f(p|1)dp \\
&= \int_{\frac{1}{2}}^{\frac{3}{4}} (8p-3)dp \\
&= \frac{1}{2}.
\end{aligned}
$$

The second line follows from $\pi = \frac{1}{2}$ and $V_0(p)f(p|0) = V_1(1-p)f(1-p|1)$.

Now let us consider a different mix of statements in the question bank – a prior of $\pi = \frac{3}{4}$. In this case, the belief associated with each signal is given by:

$$
\mu(t) = \frac{\pi(2t+1)}{\pi(2t+1)+(1-\pi)(2(1-t)+1)} = \frac{3(2t+1)}{2(2t+3)}
$$

The range of possible beliefs, in this case, is therefore $\left[\frac{1}{2}, \frac{9}{10}\right]$. The corresponding asymmetric weighted value functions are given by:

$$
\begin{aligned}
\widehat{v}_1(p) &= \frac{8p}{(3-2p)^3} - \frac{3}{(3-2p)^2} \\
\widehat{v}_0(p) &= \frac{5}{(3-2p)^2} - \frac{8p}{(3-2p)^3}.
\end{aligned}
$$

These value functions are depicted in Figure 3b. See Section A in the Appendix for details of calculations. The Simple True-False test would no longer remain optimal, for the following reason. As we see from the figure, in this case, the high prior pushes *all* beliefs above one half. So, a Simple True-False test offers *no* screening. Hence, in this case, it is appropriate for the teacher to move the belief-threshold of her True-False test *up*. But in that case, she cannot simply pass each student if and only if he gets the answer correct. That is because all the beliefs are biased towards True in this case, and choosing between passing for sure in each of the states, they would all choose True, i.e. deviate upwards. Hence, she must adjust the *passing rates* – *over-* or *under*-rewarding at least one of the chosen answers – to make the threshold type indifferent between choosing either answer, in her True-False test.

Which answer should she over/under-reward? Given the prior was already quite biased towards True, she does not want to over-reward that answer. Therefore she either over-rewards the answer False (Figure 1d) or under-rewards the answer True (Figure 1c). More concretely, she chooses some belief threshold $p > \frac{1}{2}$, and

implements it by committing to a True-False test with one of the following credit structure:

- The answer *False* is passed if it is correct and passed with probability $\left(1-\frac{1-p}{p}\right)$ if it is wrong; the answer *True* is passed if and only if it is correct.
- The answer *False* is passed if and only if it is correct; the answer *True* is passed with probability $\left(\frac{1-p}{p}\right)$ if it is correct and failed for sure if it is wrong.

The optimal belief threshold – and therefore the corresponding partial credits – can be calculated for the first case by solving the following single-dimensional optimization program:

$$\max_{p} \pi \left(\left(1-\frac{1-p}{p}\right) \int_{\underline{p}}^{\overline{p}} \widehat{v}_1(p')dp' + \left(\frac{1-p}{p}\right) \int_{p}^{\overline{p}} \widehat{v}_1(p')dp' \right) + (1-\pi) \int_{\underline{p}}^{p} \widehat{v}_0(p')dp'$$

Some algebra shows that in this particular case the above objective function is concave in p, and the optimal belief-threshold, not coincidentally, is the belief at which the weighted value functions are equal, $p = \frac{3}{4}$. We can similarly compute the teacher's optimal test of the second kind and her payoff from it. Given our parameters, the optimal test of the first kind is strictly better for her. Her payoff from it can be similarly calculated:

$$\begin{aligned} V\left(\frac{3}{4}\right) &= \frac{3}{4}\left(\left(1-\frac{\frac{1}{4}}{\frac{3}{4}}\right) \int_{\frac{1}{2}}^{\frac{9}{10}} \widehat{v}_1(p')dp' + \left(\frac{\frac{1}{4}}{\frac{3}{4}}\right) \int_{\frac{3}{4}}^{\frac{9}{10}} \widehat{v}_1(p')dp' \right) + \frac{1}{4} \int_{\frac{1}{2}}^{\frac{3}{4}} \widehat{v}_0(p')dp' \\ &= \frac{1}{4}\left(\int_{\frac{1}{2}}^{\frac{3}{4}} \left(\frac{5}{(3-2p)^2} - \frac{8p}{(3-2p)^3}\right) dp + \int_{\frac{3}{4}}^{\frac{9}{10}} \left(\frac{8p}{(3-2p)^3} - \frac{3}{(3-2p)^2}\right) dp \right) \\ &\quad + \frac{1}{2} \int_{\frac{1}{2}}^{\frac{9}{10}} \left(\frac{8p}{(3-2p)^3} - \frac{3}{(3-2p)^2}\right) dp \\ &= \frac{1}{4}. \end{aligned}$$

The optimal tests in the two variations of the above example exhibit several notable features. First, optimal tests need not exploit the full richness of the range of possible beliefs – in this example they divide that range only in two classes. Second, we observe that in the case of an unbalanced prior, a student with an extreme signal on the *opposite* of the direction in which the prior is biased might

get passed "for free", even when they are wrong. Finally, the teacher has a higher maximized payoff when the prior is balanced than when it is unbalanced. As we will show in the rest of the paper, none of these features is specific to this example, and continue to hold in a general class of environments.

4 Optimal tests

Our main result in this section is that the optimal mechanism is simple – it partitions the belief space into at most three intervals, with both a_1 and a_0 staying constant on each interval. We begin by characterizing the set of implementable mechanisms.

4.1 Implementability

The revelation principle applies in our setting. In other words, given the principal's goal of maximizing her own ex-ante payoff, it is without loss to restrict to *direct, incentive compatible* mechanisms, i.e. mechanisms where the set of messages is the set of possible beliefs P, and the pair of passing rate functions are such that the agent can never do strictly better by reporting a belief different than his own. As is standard in the literature, we assume both that the agent reports truthfully when indifferent among multiple reports, and the principal passes when indifferent.

Our first goal is to characterize the set of implementable mechanisms. Since there are no transfers, this set is defined only by the agent's incentive compatibility constraints:

$$pa_1(p) + (1-p)a_0(p) \geq pa_1(p') + (1-p)a_0(p'), \forall p, p' \in P \tag{IC}$$

The indirect utility of an agent with belief p is given by $U(p) := pa_1(p) + (1-p)a_0(p) = a_0(p) + p(a_1(p) - a_0(p))$. Note that this expression is identical to that for the agent's indirect utility in the standard monopolistic screening problem, with $q := a_1 - a_0$[15] playing the role of the "allocation", and $-a_0$ playing that of the "transfer". Clearly, $q \in [-1, 1]$. Any mechanism (a_1, a_0) can therefore isomorphically be described by (U, q) where U and q are as defined above.

With the above transformation, direct application of results from the standard monopolistic screening setting allows us the following simplification.

Lemma 1. *A mechanism (U, q) is incentive compatible if and only if q is non-decreasing, and the agent's indirect utility is given by:*

[15] Throughout the paper we use the convention, that when the name of a function is used without its argument, it denotes the function as a point in its respective vector space.

$$U(p) = U(\underline{p}) + \int_{\underline{p}}^{p} q\left(p'\right) dp' \tag{1}$$

for all p.

Proof. Standard. See appendix. □

Using the fact that $U(p) = pa_1(p) + (1-p)a_0(p)$ in conjunction with Lemma 1 gives us the following "integral formulas":

$$\begin{aligned} a_0(p) &= \left(U(\underline{p}) + \int_{\underline{p}}^{p} q\left(p'\right) dp'\right) - pq(p) \\ a_1(p) &= \left(U(\underline{p}) + \int_{\underline{p}}^{p} q\left(p'\right) dp'\right) + (1-p)q(p) \end{aligned} \tag{Integral Formulas}$$

4.2 Optimal tests are coarse

We solve the principal's constrained optimization problem, which – given her general linear objective function – gives us the extreme points of the set of implementable mechanisms described above.

The principal's problem is as follows:

$$\max_{a_1, a_0 \in [0,1]^P} \quad \pi \int_p v_1(p) a_1(p) dF(p|1) + (1-\pi) \int_p v_0(p) a_0(p) dF(p|0) \tag{1}$$

$$\text{s.t. } a_1, a_0 \in [0,1]. \tag{Feas}$$

$$pa_1(p) + (1-p)a_0(p) \geq pa_1\left(p'\right) + (1-p)a_0\left(p'\right), \forall p, p' \in P \tag{IC}$$

In the rest of the section, we first present and discuss the results, providing some intuition, and then move on to the proof sketch.

Theorem 1 (Optimal tests are simple)**.** *The optimal mechanism* (a_1, a_0)*, which solves* (1)*, consists of step functions with at most two steps, where the* a_1 *and* a_0 *change at the same belief(s).*

In Section B in the appendix we provide a more detailed characterization of the exact form of the optimal mechanism, specifying formulas for the thresholds of these steps, and associated passing probabilities (Theorem B.1).

In order to provide some intuition for why the optimal mechanisms take such a simple form, we start from the first best benchmark and construct the optimal mechanisms through gradual adjustments. As we saw in the Example (Section 3), if the principal faced no incentive constraints, she would want to pass the agent with a given belief if and only if her *value* from passing him given the true answer is

positive. This implies that under the optimal unconstrained mechanism (a_1 and a_0) take values only in $\{0, 1\}$, are potentially non-monotone, and change at potentially different belief thresholds.

Incentive compatibility can be restored to this ideal mechanism if and only if adjustments are made to it to ensure the following: (i) coincidence of thresholds (ii) monotonicity, and (iii) indifference at the thresholds.[16] Below we discuss why this holds and show how making these adjustments to the first-best mechanism immediately leads to the class of optimal mechanisms. An agent cares only about his weighted average probability of passing if the correct answer is 1 and 0, with his belief dictating the weights. The higher the belief, the more (respectively less) the agent cares about his probability of being passed in state 1 (respectively 0). Consider two agents A and B such that A considers answer 1 to be more probable than B. Since both care about the interim probability of being passed, if the passing probability of A in one of the states is more than B, then his passing probability in the other must be less than B, to ensure truthful reporting. Within the class of threshold mechanisms, this would mean the thresholds of rising of a_1 and falling of a_0 must be coincident.

We also need monotonicity, for the following reasons. Note that if B is passed with a higher probability in state 1 than A, B's passing probability in state 0 must be low enough so that A does not find it beneficial to masquerade as B. But because B cares about state 0 *more* than A, this would entice B to misreport as A. Therefore A's passing probability in state 1 (respectively, 0) must be more (respectively, less) than B's, i.e. both a_1 and a_0 must be monotonic.

Suppose the principal faces just the above two restrictions - common thresholds and monotonicity - and no other. In this case she would like to use a single-threshold mechanism where both passing probabilities undergo a dramatic change at the threshold - a_1 from 0 to 1 and a_0 from 1 to 0.[17] But unless this threshold is one-half, such a mechanism is not incentive compatible - if types close to the threshold type think the correct answer is more likely to be 1, between getting passed for sure when the correct answer is 1 vs 0, they choose the former, *misreporting* if they are below the threshold.

There are two types of adjustments the principal can make to her favorite common threshold bang-bang mechanism described above - on the extensive and intensive margins - to achieve full incentive compatibility, fulfilling condition (iii). She makes adjustments on the *extensive* margin when she changes *who* to pass - by adjusting the threshold. Conversely, adjustments on the *intensive* margin consist of changing the *rate* at which to pass, given a threshold. How she trades off between

[16] Formally, a piece-wise constant mechanism (a_1, a_0) is incentive compatible if and only if it satisfies those three conditions.

[17] This follows intuitively from the discussion in the previous paragraphs, but can be shown formally.

these two types of adjustments is determined by the specifics of the setting.

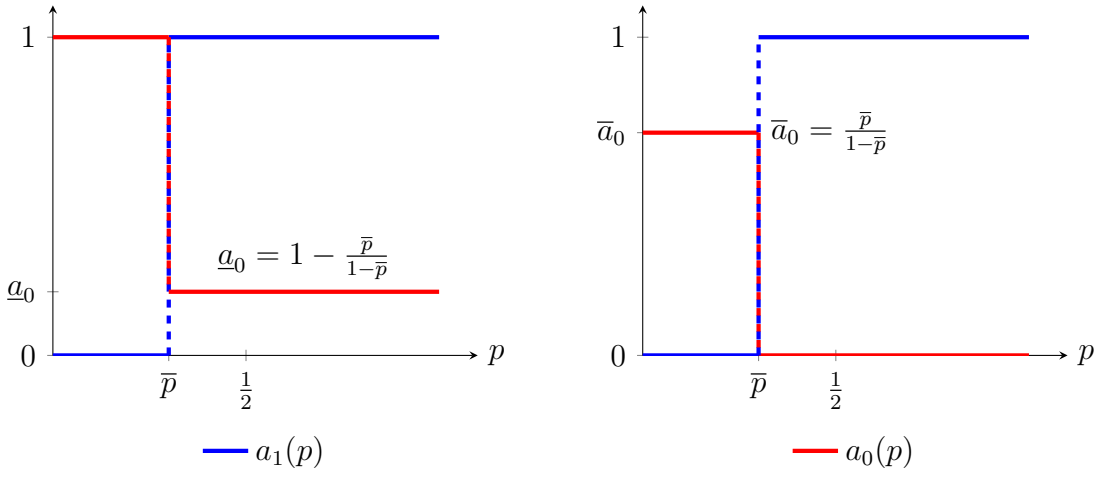
Each combination of adjustments gives rise to a different class of optimal mechanisms, which solve (1). On the one hand, the principal can adjust *only* on the extensive margin - adjusting the common threshold to one half without adjusting the passing rates. The resulting mechanism, depicted in Figure 1a and called the *simple True-False* test, arises as the optimal mechanism generically in a broad class of settings (See Section 5.2.1). On the other hand, if she does choose to adjust on the intensive margin, in general the optimal positioning of thresholds also changes as a result, leading to an adjustment on the extensive margin as well. Loosely speaking, she wants these adjustments to be to be minimal, so she adjusts the the passing rate only in one of the states and on one of the sides of the optimal threshold an optimizes the threshold accordingly. Each of the four classes of resulting mechanisms (Figures 1c-1d and their flipped versions around the x axis) can arise optimally, as shown in Section 5.2.1.

4.2.1 Single threshold tests: "Pass-if-correct" and "Fail-if-incorrect"

In this subsection we describe in more details the single-threshold tests which arise from the above adjustments, as these would prove key to subsequent results in the rest of the paper. Let us take a common threshold bang-bang test with a threshold below one half. The threshold type is not indifferent – he strictly prefers reporting the "low message" over the "high message" because his belief is biased towards zero. Therefore, there are two ways we can make this test incentive compatible – by making the high message more lucrative or the low message less so. This can only be done by increasing the passing rate in state 0 for the high message or decreasing it for the low message. We call the two types of tests which arise *pass-if-correct* and *fail-if-incorrect* tests, respectively.

The formal definitions of *pass-if-correct* and *fail-if-incorrect* tests are as follows. Let p^ω denote the agent's belief that the state is ω, i.e. $p^1 = p$ and $p^0 = 1 - p$. We call a test *pass-if-correct* if and only if it is given by $(a_\omega(p^\omega) = 1(p^\omega \geq p_0), a_{\omega^c}(p^\omega) = 1(p^\omega \leq p_0) + p^* 1(p^\omega \geq p_0)$ for some $\omega \in \{0,1\}, p^* \in (0,1)$. Similarly we call it a *fail-if-incorrect* test if and only if it is given by $(a_\omega(p^\omega) = 1(p^\omega \geq p_0), a_{\omega^c}(p^\omega) = p^* 1(p^\omega \leq p_0))$ for some $\omega \in \{0,1\}, p^* \in (0,1)$.

The names derive from the fact that any single threshold tests can be indirectly implemented by a "pick the correct answer" test with two options, with credits specified for each combination of the correct and given answers, as depicted below. By correctly specifying the credits, any desired threshold can be implemented.



(a) Pass-if-correct (b) Fail-if-incorrect

Figure 4: Direct and corresponding indirect implementations of Pass-if-correct and Fail-if-incorrect tests

		Given Answer	
		T	F
Correct Answer	T	100%	$1 - \frac{\overline{p}}{(1-\overline{p})}$
	F	0%	100%

(a) Pass-if-correct

		Given Answer	
		T	F
Correct Answer	T	$\frac{\overline{p}}{(1-\overline{p})}$	0%
	F	0%	100%

(b) Fail-if-incorrect

We wrap up this section with an overview of our proof strategy.

4.2.2 Proof sketch

Using the (Integral Formulas) derived in Section 4.1, we can express the principal's ex-ante value in terms of her "virtual value" function χ as follows:

$$\mathbb{V}(a_1, a_0) = U(\underline{p})\mathbb{E}v + \int_{\underline{p}}^{\overline{p}} \chi(p)q(p)dp \tag{2}$$

We abstract from the exact expression for $\chi(p)$ in this section, as it is not relevant to the proof sketch. We introduce it in Section 5.1.1, where we formulate a standard "regularity" condition in terms of our virtual value. See Section D.1.3 in the Appendix for the derivation of (2) and Section C.1 for its economic significance, as well as that of q ("knowledge premium") introduced above.

Unlike in a standard monopolistic screening problem, our principal faces *feasibility* constraints, namely the fact that $a_1, a_0 \in [0, 1]$. But due to their monotonicity, as seen from (Integral Formulas), it is sufficient to ensure that these feasibility constraints are satisfied at the extremes, namely, that $a_0(\underline{p}) \leq 1$, $a_0(\overline{p}) \geq 0$, $a_1(\underline{p}) \geq 0$

and $a_1(\overline{p}) \leq 1$. Combining these with the IC constraints we obtain a reduced-form of the principal's problem, as given by the following lemma.

Lemma 2. *The principal's optimal mechanism is given by the solution to the following problem where a_1 and a_0 are as given by (Integral Formulas) and $p_0 := \sup\{p : q(p) \leq 0\}$.*

$$\max_{U(\underline{p}),q(.)} U(\underline{p})\mathbb{E}v + \int_{\underline{p}}^{\overline{p}} \chi(p)q(p)dp \tag{Problem}$$

$$\begin{aligned}
s.t.\ & q \in [-1,1], q \text{ non - decreasing} && \text{(MON)}\\
& a_0(\underline{p}) \leq 1, && \text{(F1)}\\
& a_0(\overline{p}) \geq 0, && \text{(F2)}\\
& a_1(\underline{p}) \geq 0, && \text{(F3)}\\
& a_1(\overline{p}) \leq 1, && \text{(F4)}\\
& \text{where } a_1 \text{ and } a_0 \text{ are given by (Integral Formulas).} && \text{(Integral)}
\end{aligned}$$

An overview of the rest of the proof is as follows. We show that *each* bound - 1 and 0 - must be attained at least once. Moreover, unless there is no screening - i.e. the optimal mechanism is constant with respect to agents' beliefs - each extreme of the belief space must have at least one of a_1 and a_0 attaining its respective bound there. Combining these two insights shows that whenever there is screening at the optimum, the above program reduces to one constrained by (MON) and a single linear equality. We know the extreme points of the convex set of q's described by (MON) are step functions (including degenerate ones). So those of the convex subset of it described by imposing the additional linear constraint are obtained by taking convex combination of at most two of them. Theorem 1 follows.

5 Rewarding ignorance and penalizing correct answers

In this section we delve deeper into the two most noteworthy features of the optimal tests – that they might reward admission of ignorance (Figure 1b) and that they may fail agents giving the correct answer or pass those giving the wrong one (Figures 1c and 1d). We provide conditions when each of these cases arises. Further, our comparative statics analysis uncovers a link between "obviousness" of an answer and the rate at which it is failed in spite of being correct or passed in spite of being wrong.

For the sake of interpretability of the aforementioned results, first, we concretize our model further, as described below.

5.1 A microfoundation

Suppose the value that the agent generates for the principal when passed is v, to be interpreted as his *quality*. We assume higher quality agents are more effective at learning (to be made precise shortly). Hence, sometimes we also refer to v as the agent's *learning type*. Let $v \in \mathcal{V} := [\underline{v}, \overline{v}], \underline{v} < 0 < \overline{v}$. Further, let v be distributed according to some commonly known probability measure $\nu \in \Delta\mathcal{V}$. Paralleling (Non-triviality), we assume $\nu(\{v < 0\}) > 0$ and $\nu(\{v > 0\}) > 0$, i.e. both positive and negative qualities occur with positive probability. Let $V_0 := \mathbb{E}_\nu v$ denote the a priori expected quality. Quality is unobserved by both the principal and the agent.[18] Quality and the state ω are independent.

The agent's learning technology is as follows. There is a signal space $T := [0, 1]$. We denote the typical signal realization by $t \in T$. The agent's signal is drawn from some known distribution over T for each each state ω and quality v, which is assumed to have a density, denoted by $f(\cdot|\omega, v)$, $(\omega, v) \in \{0, 1\} \times \mathcal{V}$. Going forward, we refer to the pair of signal densities $(f(\cdot|1, v), f(\cdot|0, v))$ as the *signal structure* of the agent of quality v. We impose a number of assumptions on the family of signal structures $\{(f(\cdot|1, v), f(\cdot|0, v))\}_{v\in\mathcal{V}}$, as detailed below.

Assumption 1. *Suppose the following holds.*

- ***Monotonicity:*** *$f(t|1, v)$ (respectively $f(t|0, v)$) is increasing (respectively decreasing) in t for all v.*
- ***MLRP:*** *The family of signal structures $\{(f(\cdot|1, v), f(\cdot|0, v))\}_{v\in\mathcal{V}}$ satisfies the Monotone Likelihood Ratio Property (MLRP). That is, $\frac{f(t|1,v)}{f(t|1,v')}$ is increasing and $\frac{f(t|0,v)}{f(t|0,v')}$ decreasing in t for all $v > v'$.*
- ***C^2:*** *$f(\cdot|\omega, v)$ is twice continuously differentiable in t for all ω, v.*
- ***Boundedness:*** *$\sup_{t,e,\omega} f(t|\omega, v) < \infty$ and $\sup_{t,e,\omega} f'(t|\omega, e) < \infty$.*

We call the signal structures *symmetric* if $f(t|0, v) = f(1 - t|1, v)$ for all v.

Since the agent does not observe his own quality, the signal distribution from which his observed signals are drawn – denoted by $\{\overline{f}_\omega\}_{\omega\in\{0,1\}}$ – is the *mean* signal distribution of the various learning types:

$$\overline{f}_\omega(t) = \int_e f(t|\omega, v)d\nu(v), \ \forall\, t, \omega$$

[18] In Section 7.1 we use the alternative modeling assumption that the agents do observe their quality. The main features of the optimal test remain unaltered.

He forms his beliefs about the state using Bayes rule using the above distribution. Using $\mu : [0,1] \to [0,1]$ to denote the mapping between signal and belief spaces, we have using Bayes rule, for a given prior π,

$$p = \mu(t) := \frac{\overline{f}_1(t)\pi}{\overline{f}_1(t)\pi + \overline{f}_0(t)(1-\pi)} \tag{1}$$

Note that by our assumptions on $\{(f(\cdot|\omega, v)\}_{(\omega,v)\in\{0,1\}\times\mathcal{V}}$, $\mu(t)$ is strictly increasing for any given π. Hence there is a one-to-one map between signals and beliefs, i.e. μ^{-1} exists. Let $(\widehat{a}_1, \widehat{a}_0)$ denote the mechanism with the *signal* space as its domain, with $\widehat{a}_\omega : [0,1] \to [0,1]$ denoting the passing probability as a function of the reported signal, in state ω. Hence, $\widehat{a}_\omega = a_\omega \circ \mu^{-1}, \omega \in \{0,1\}$.[19]

In this specialization of our general model from Section 2, while choosing the mechanism, the principal exploits the correlation between the informativeness of the signal structures and agent quality to maximize the expected quality conditional on passing. With some algebra, her payoff can be described in terms of "summarized" *value functions* for each state – m_1 and m_0 – as given below. See Section D.1.1 in the Appendix for exact expressions for m_1 and m_0 in terms of the family of signal structures $\{(f(\cdot|\omega, v)\}_{(\omega,v)\in\{0,1\}\times\mathcal{V}}$.

$$V(\widehat{a}_1, \widehat{a}_0) := \left[\pi\left(\int_0^1 \widehat{a}_1(t)m_1(t)dt\right) + (1-\pi)\left(\int_0^1 \widehat{a}_0(t)m_0(t)dt\right)\right] \tag{Principal's Value}$$

5.1.1 Regularity

In this subsection, we provide a regularity condition which ensures that the optimal tests feature a single threshold. While none of our subsequent analyses – such as comparative statics and endogenizing the prior – relies on a restriction to single-threshold tests, we use the "regular" class of problems to offer simplified versions of some of the results.

$$\chi(p) := \int_p^1 \pi v\left(p'\right) dp' - pv(p) + \pi V_1(p) dF(p \mid 1), ; \text{ where}$$
$$v(p) = \pi V_1(p) dF(p \mid 1) + (1-\pi) V_0(p) dF(p \mid 0)$$

It is well-known that in the standard monopolistic screening problem, such a regularity condition is equivalent to concavity of the monopolist's revenue in the price. In a similar vein, in the rest of the subsection we show that within the microfounded model introduced above, our regularity condition is equivalent to the concavity of

[19] Note that μ —and therefore the mapping between $\widehat{a}_\omega$ and $\widehat{a}_\omega$ – depends on the prior π.

the principal's value as a function of the threshold of a single-threshold test. But first, we need to unavoidably define two new objects – the agent's *normalized likelihood ratio* and the principal's value from each single threshold test as a function of the threshold. These objects are key to our main result in this subsection as well as that in Section 5.3 – characterization of when the optimal test rewards ignorance, i.e. features two thresholds.

We define the *normalized likelihood ratio* (hereafter, NLR) function for state ω, $\phi_\omega : T \to \mathbb{R}_+$, $\omega \in \{0, 1\}$, as the likelihood ratio of each signal for state ω, normalized by the prior likelihood ratio of the same state. Specifically, for each $t \in T$,

$$\phi_1(t) := \frac{\frac{\mu(t)}{1-\mu(t)}}{\frac{\pi}{1-\pi}} = \frac{\overline{f}_1(t)}{\overline{f}_0(t)}, \ \phi_0(t) = \frac{1}{\phi_1(t)} \tag{2}$$

Under any single-threshold optimal test with a belief threshold other than one half, the passing rate in only one of the states is distorted on the intensive margin. For any such test, let us call this the *distorted state* and the other state – the one where the passing rate is bang-bang – the *undistorted state*.

Define $v_\omega : \phi_\omega(T) \to \mathbb{R}$ as the function which maps NLR's to the principal's maximized value from a single-threshold test with that NLR-threshold and undistorted state ω. For example, *given* a signal threshold t and undistorted state 1, if the corresponding fail-if-incorrect test does better for the principal than the pass-if-correct test,[20] letting $s_1 = \phi_1(t)$ and using the expression for (Principal's Value), we have,

$$\begin{aligned} V_1(s_1) &= \pi \int_{\phi_1^{-1}(s_1)}^{1} m_1(t)dt + (1-\pi) \times \left(\frac{\pi}{1-\pi}\right) \times s_1 \int_{0}^{\phi_1^{-1}(s_1)} m_0(t)dt \\ &= \pi \left(\int_{\phi_1^{-1}(s_1)}^{1} m_1(t)dt + s_1 \int_{0}^{\phi_1^{-1}(s_1)} m_0(t)dt \right). \end{aligned}$$

Alternatively, if the pass-if-correct test does better, we have,

[20] In Proposition 2 in Section 5.2.1 we show that the fail-if-incorrect (respectively, pass-if-correct) test does strictly better if and only if $V_0 < 0$ (respectively, $V_0 < 0$). We postpone that discussion till the next section, as it is not relevant to the results presented here.

$$V_1(s_1) = \pi \int_{\phi_1^{-1}(s_1)}^{1} m_1(t)dt + (1-\pi)\left[\int_0^{\phi_1^{-1}(s_1)} m_0(t)dt + \left(1 - \left(\frac{\pi}{1-\pi}\right)s_1\right)\int_{\phi_1^{-1}(s_1)}^{1} m_0(t)dt\right]$$
$$= \pi \int_{\phi_1^{-1}(s_1)}^{1} \left(m_1(t) - s_1 m_0(t)\right) dt + (1-\pi)V_0$$

In general, $V_1(s_1)$ is the maximum of the above two expressions. This can be compactly written as:

$$V_1(s_1) = \pi\left(s_1\left(\int_0^{\phi_1^{-1}(s_1)} m_0(t)dt - v_0^+\right) + \int_{\phi_1^{-1}(s_1)}^{1} m_1(t)dt\right) + (1-\pi)v_0^+$$

Similarly, we can derive the expression for $V_0(s_0), s_0 \in \phi_0(T)$. See Section D.6 in the Appendix for details.

It can be shown that increasingness of the virtual value – our regularity condition – is equivalent to concavity of the value functions defined above, as formalized below.

Proposition 1. *The following are equivalent.*

- V_1 *is concave.*
- V_0 *is concave.*
- *The problem is regular.*

The intuition is as follows. Why does increasingness of the virtual value lead to sufficiency of single-threshold mechanisms in screening problems? It is because it ensures concavity of the principal's value from a single-threshold mechanism, as a function of the threshold. In our case, this is her value from using a single-threshold bang-bang test (not necessarily incentive compatible) as a function of the belief-threshold. This ensures the principal cannot gain from offering a lottery over two such tests, which is what a two-threshold test is. Our proof shows, that translated to the microfounded version of our model, such a lottery is essentially taken over two single-threshold incentive compatible – but not necessarily feasible (because passing rates can go beyond 1) – tests. Naturally, if the principal's value (V_1 and V_0) is concave in the threshold, she cannot gain from such a lottery.

It can also be easily shown that the regularity condition does not depend on the prior – a fact we would exploit in our comparative statics analysis going forward.

Fact 1. *The problem is regular for prior π if and only if it is regular for all other priors $\pi' \neq \pi$.*

5.2 Over-rewarding counterintuitive answers

5.2.1 Comparative statics

In this section we analyze how the optimal test varies with the prior. We show that the more extreme the prior, the greater the distortion on the intensive margin.[21] Moreover, if the prior is moderate, the principal cannot benefit from such distortions, and chooses the simple True-False test instead. Whenever she has to distort on the intensive margin, it gives rise to *asymmetric* rewards for the correct answer in the two states. Our main result in this section is that, such a scheme always ends up rewarding the *counterintuitive* answer – an answer which is a priori sufficiently unlikely – more.

While none of the results in this section require regularity, for drawing the above insights we focus on regular settings. Under regularity, there are only the four types of single-threshold tests (Figures 1c-1d and their flipped versions around the x axis) which can be optimal. As described in section 4.2.1, they can further be divided into two categories – those which always fail incorrect answers ("fail-if-incorrect") and those which always pass correct answers ("pass-if-correct").

When does each of these two types of tests arise as optimal?

It turns out, that depends only on the a priori expected quality, as described below.

Cherry-picking and lemon-dropping markets. Note that if the principal could not screen, she would pass the agent if and only if the a priori expected quality is positive, i.e. $V_0 \geq 0$. Hence we call markets with $V_0 < 0$ (respectively $V_0 \geq 0$) *cherry-picking* (respectively *lemon-dropping*) markets, because the principal screens to "pick the cherries", i.e. identify the acceptable agents (respectively "drop the lemons", i.e. weed out unacceptable agents).

The lemma below captures the fact that the type of test – pass-if-correct or fail-if-incorrect – depends *only* on the type of market.

Proposition 2. *Pass-if-correct (respectively fail-if-incorrect) tests are optimal only if the market is lemon-dropping (respectively cherry-picking).*

The intuition behind the above insight is straightforward, once we note that for a fixed threshold, the pass-if-correct test is identical to the fail-if-incorrect test, except its passing rate in the distorted state[22] is pushed upwards uniformly by a constant

[21] Recall from Section 4.2 that we call adjustments in the *rate* of passing (as opposed to the *threshold* of passing), away from the principal's favorite levels of 1 or 0 – whichever is better – a distortion on the *intensive margin*.

[22] Under any single-threshold optimal test with a belief threshold other than one half, the passing rate in only one of the states is distorted on the intensive margin. Recall the forllowing terminology

amount for all beliefs. The jump of the passing rate in the distorted state between the correct and wrong answers is fixed by incentive compatibility[23], regardless of the type of test. Hence for a fixed threshold, the only thing the principal is free to choose is whether to offer a positive baseline passing rate in one of the states. By definition, she gains (respectively, loses) from such default passing if and only if the market is lemon-dropping (respectively, cherry-picking).

Now we are ready to present the main result of this section.

Theorem 2.A (Variation of the optimal test with the prior)**.** *For any problem $\mathcal{P}$ there exist $0 < \underline{\pi} < \overline{\pi} < 1$ such that the type of the optimal test is always pass, never pass or given by the following:*

	$\pi < \underline{\pi}$	$\pi \in [\underline{\pi}, \overline{\pi}]$	$\pi > \overline{\pi}$
Cherry-picking, $V_0 < 0$	Fail-if-incorrect, penalize False	Simple T-F or T-F-U, Simple T-F if regular	Fail-if-incorrect, penalize True
Lemon-dropping, $V_0 \geq 0$	Pass-if-correct, bonus for True		Pass-if-correct, bonus for False

Table 3: Optimal tests across markets and priors

Moreover, if the signal structures are symmetric, $\underline{\pi} + \overline{\pi} = 1$.

The intuition for the above result is as follows. Let us consider only the cherry-picking case, to simplify terminology. The belief-threshold in a single-threshold test represents a *bar* of passing – it is the lowest belief *for* a state (i.e., an answer) which could pass when that state realizes (i.e., when that is the correct answer).[24] Its choice constitutes a trade-off – the higher it is in one state, the lower it is in the other. In other words, the more difficult the test is if the correct answer were 1, the easier it is if it were 0. A belief threshold of one half offers the same bar of passing in both states.[25] As we would show in the next section, a two threshold test is a lottery over two single-threshold tests – one with a belief threshold above one half and the other below, i.e. one with a bar which is high for state 1 and hence low for state 0, and the other, the reverse.

we introduced in Section 5.1.1: for any such test, we called this state the *distorted state* and the other state – the one where the passing rate is bang-bang – the *undistorted state*.

[23]Specifically, $\left(\frac{p_0}{1-p_0}\right)$ for a threshold belief $p_0 < \frac{1}{2}$, for example, and conversely, its reciprocal, when $p_0 > \frac{1}{2}$.

[24]In the lemon-dropping case, this is the highest belief for that state which could fail if that state realizes.

[25]Note that the *bar* of passing is not the same as the belief-threshold of passing. The former depends on the state. It is equal to the latter only in state 1, and is its complement in state 0. Hence, while all single-threshold tests have the same belief-threshold of passing in both states, only the one with the threshold of one half has the same bar.

If the prior is too biased towards one of the states, resolving the aforementioned trade-off is straightforward for the principal – she wants the bar for passing to be high (i.e. greater than one half)[25] in *that* state. The reason is as follows. Suppose the prior is very high. In this case, *so small* a part of the population would have a belief below one half, that a belief-threshold at or below it would lump nearly all of the population together, offering very little screening. Hence, in this case, the principal would benefit neither from mixing a high-bar test with a low-bar one, nor from offering the same bar in both the states with a Simple T-F. Conversely, when the prior is moderate, either of these compromises would make the her better off than having a high bar only in one of the states.

Formalizing the above notion, going forward, given a problem $\mathcal{P}$ we call corresponding priors $\pi \in [\underline{\pi}, \overline{\pi}]$ *moderate* and those $\notin [\underline{\pi}, \overline{\pi}]$, *extreme* priors. An extreme prior π is said to *grow more moderate* if it changes towards the nearest moderate prior, i.e. when it increases for $\pi < \underline{\pi}$ and when it decreases for $\pi > \overline{\pi}$.

We call the answer-option mirroring the a priori likely state the *obvious* answer and the other one the *counterintuitive* answer. As we saw above, whenever the prior is extreme, the principal must move the optimal threshold away from the balanced threshold of one half, leading to a distortion – a partial reward for a *wrong* answer or a penalty for the *correct* answer, in terms of the indirect implementations of the optimal tests – in one of the states. Which answer does such a distortion benefit or hurt. As Theorem 2.A details, it is the counterintuitive answer which always benefits, compared to the obvious answer, whenever the prior is extreme. This is formalized below.

Corollary 1. *When the prior is extreme either there is a bonus for the counterintuitive answer or a penalty for the obvious answer.*

Theorem 2.A tells us how the first trade-off faced by the principal - whether to fine-tune the threshold - is resolved. The second trade-off she faces is *how much* to fine-tune the threshold, when such fine-tuning is warranted, i.e. the prior is extreme. This is elaborated by Theorem 2.B below.

Theorem 2.B (Variation of the optimal test with the prior)**.**

- *When the prior is extreme, as it grows more moderate, the signal threshold of the optimal test remains constant, its belief threshold grows more moderate and the distortion in the allocation in the a priori obvious state reduces.*

- *Suppose the optimal test has two thresholds for some moderate prior $\pi \in (\underline{\pi}, \overline{\pi})$, with signal thresholds t_1 and t_2, $t_1 > t_2$. Then:*
  - *It has the same two signal thresholds for all priors $\pi \in \left(\frac{1}{\phi_1(t_1)+1}, \frac{1}{\phi_1(t_2)+1}\right)$.*
  - *As the prior increases from $\frac{1}{\phi_1(t_1)+1}$ to $\frac{1}{\phi_1(t_2)+1}$, $\widehat{a}_1(t)$ increases from 0 to 1 and $\widehat{a}_0(t)$ decreases from 1 to 0, for all $t \in (t_2, t_1)$.*

– *The optimal test for* $\pi = \frac{1}{\phi_1(t_i)+1}$ *is the Simple T-F with signal threshold* t_i, $i \in \{1, 2\}$.

- *Finally, if the optimal test is the Simple T-F for an open interval of moderate priors, as the prior increases within this interval, the signal threshold decreases.*

The intuition for the above result in the case of an extreme prior is as follows. Suppose $\pi > \underline{\pi}$ and the signal threshold for the optimal test is t. As the prior grows more extreme, the belief associated with t also grows more extreme – further away from one half. So, the distortion on the intensive margin under the optimal test increases. The principal can attempt to rectify the associated loss by making adjustments on the extensive margin, i.e. by making the signal threshold moderate. In this case, that would mean decreasing it. However, this would *lower the bar* for passing in the a priori likely state 1, passing beliefs which are too inaccurate and thereby hurting the principal's payoff in that state. This loss also becomes more costly to her as the prior grows more extreme, since a priori likely state 1 becomes more likely.

A similar pair of forces come into consideration for the principal, when she decides whether to vary the signal- thresholds of a two-threshold test with the prior. A two-threshold test is a lottery over two single-threshold bang-bang tests, one of which is preferred by the principal over the other. Suppose that is the test with the signal-threshold t_2, as described in the statement of Theorem 2.B. As the prior increases within the range $\left(\frac{1}{\phi_1(t_1)+1}, \frac{1}{\phi_1(t_2)+1}\right)$, the distortion on the intensive margin, vis-a-vis the principal's favorite bang-bang test with threshold t_2, increases. She could counteract this by bringing t_1 closer to t_2, but that would increase the distortion on the extensive margin (lower the bar of passing).

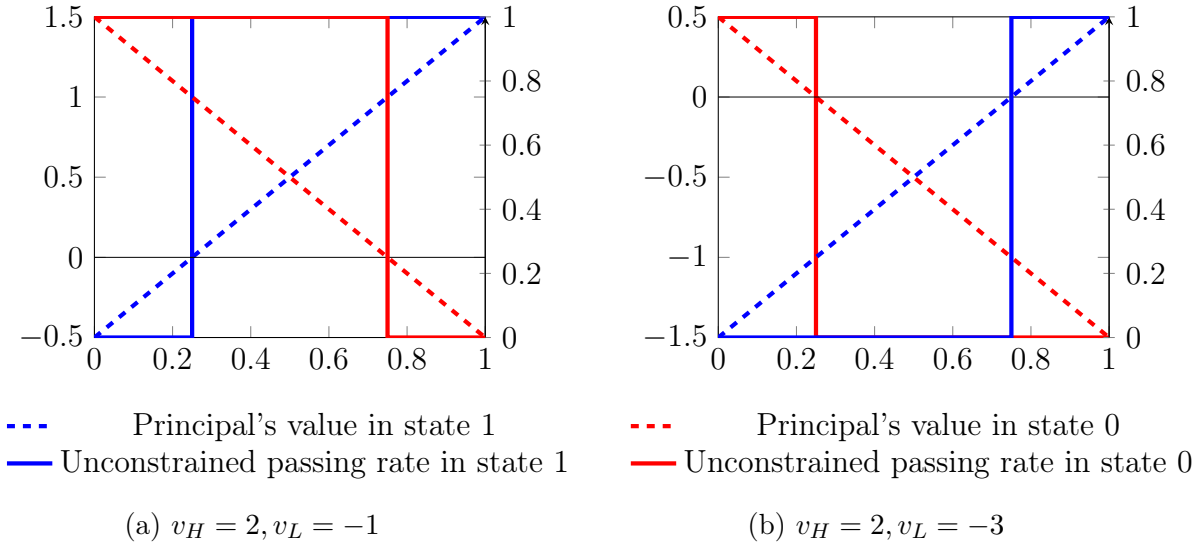
In both of the above cases of extreme and moderate priors, these two forces – those of minimizing the distortion on the intensive and extensive margins – exactly cancel each other out, leaving the principal's optimal signal-threshold(s) invariant to the prior within a range. It is important to note that this exact cancellation happens precisely because the principal and the agent share a prior. As we show in Section 7.2.1, this would not be the case if the priors are different, in which case the signal threshold(s) *would* vary with (each) prior even within the aforementioned ranges.

5.2.2 Distortions

In this section we analyze welfare effects of the distortions introduced by agency costs. Turns out, the theme of counterintuitive answers being over-rewarded and obvious answers being over-penalized carries over to welfare effects as well. Our main insight is that agents with strong convictions about the counterintuitive answer are

better off and those with strong convictions about the obvious answer are worse off, compared to the complete information benchmark.

The principal uses Bayes rule to update her beliefs about quality, based on reports of signals, which, in turn drive her valuation of each signal. By (Principal's Value), the complete information – or *first best* – mechanism is given by $a_\omega^{fb}(t) = 1(m_\omega(t) \geq 0), \omega \in \{0,1\}$. An example is illustrated in the Figure 5.



(a) $v_H = 2, v_L = -1$ (b) $v_H = 2, v_L = -3$

Figure 5: First best tests

Example: The principal's value from passing type e is v_e, $e \in \{H, L\}$. With prior π, the principal's expected payoff from a mechanism $\widehat{a}_1, \widehat{a}_0 : T \to [0,1]$ is given by (Principal's Value), with $m_1(t) = 2t - \alpha, m_0(t) = 2(1-t) - \alpha$, where $\alpha = \frac{|v_L|}{v_H}$.

Several features of the first best test in this example are notable. First, extreme beliefs are passed if and only if they are "correct", i.e. sufficiently close to the true state. Second, moderate beliefs are either passed or failed, regardless of their correctness. And finally, the test does not depend on the prior beliefs. None of these features are specific to this example however, as the below proposition shows.

Theorem 3.A. *For any problem $\mathcal{P}$, unless the first-best test is to always or never pass, there exist $t_0^{fb}, t_1^{fb} \in (0,1)$ such the first-best test $a_0^{fb}, a_1^{fb} : T \to [0,1]$ satisfies the following properties:*

- $a_0^{fb}(t) = 1(t \leq t_0^{fb}), a_1^{fb}(t) = 1(t \geq t_1^{fb})$.
- *Types $t \in [\min\{t_0^{fb}, t_1^{fb}\}, \max\{t_0^{fb}, t_1^{fb}\}]$ are either passed or failed in both states.*
- *The test does not depend on the prior.*

Comparing the first best test with the structure of the optimal constrained test established in previous sections, we note the following.

Theorem 3.B (Distortions). *For any regular problem there exists* $0 < \underline{t} \leq \bar{t} < 1$ *such that the distortions for moderate and high priors are given in the following table, where* $\underline{\pi}$ *and* $\overline{\pi}$ *are as defined in Theorem 2.A.*

		Prior		
		$\pi > \overline{\pi}$		$\overline{\pi} < \pi < \underline{\pi}$
		Cherry Picking, $V_0 < 0$	Lemon Dropping, $V_0 \geq 0$	
Agent Signal	$[0, \underline{t}]$	No Distortion	Better Off	No Distortion
	$[\underline{t}, \bar{t}]$	Ambiguous		
	$[\bar{t}, 1]$	Worse Off	No Distortion	No Distortion

Table 4: Strong beliefs are over-rewarded when they are counterintuitive

Analogous results hold for $\pi > \overline{\pi}$*, with the rows reversed.*

The reasoning behind the above result is straightforward. For moderate priors, by Theorem 2.A, there is no distortion on the intensive margin. This leads to undistorted allocations for extreme types. For extreme priors, we know from Theorem 2.A that the beliefs which lean towards the obvious state are passed at a rate which is weakly distorted downwards in that state (strictly if $V_0 < 0$). The converse holds for beliefs leaning towards the counterintuitive state. This explains the direction of distortion for extreme priors.

For intermediate types, $t \in [\underline{t}, \bar{t}]$, the sign of the distortion depends on the prior ν. The details for those types, along with the proof are presented in the Appendix.

Comparing Theorem 3.B with analogous results for standard screening models offers two key insights. First, it shows that unlike in those models, where screening makes all agent types weakly better off vis a vis the full information benchmark, in our model screening can both help and hurt the agent. This is because the objectives of the principal and the agent are not fully opposed in our model, unlike in standard screening models where the transfer component typically affects the two parties in directly opposed ways. Hence, information rent can be *negative* in our model – a common feature of models of mechanism design without transfer (e.g. Ben-Porath et al. (2014)).

Secondly, it further underscores the force in the model which favors agents with strong counterintuitive beliefs. A broad class of screening models feature the well-known no-distortion-at-the-top property (Mussa and Rosen (1978),Myerson (1981), Rochet and Choné (1998),Rochet and Stole (2002) etc.). In these models typically only "downward" incentive constraints bind. Therefore the only reason the allocation of a type can be distorted downwards is to discourage impersonation by higher types. Consequently the "highest" type's allocation is undistorted because there are no higher types which can impersonate it. Even in our model, the incentive

constraints bind only in one direction – from counterintuitive towards obvious. When the passing rate of the obvious answer is distorted downwards it is to discourage the worst – i.e. most moderate – among those giving the latter answer, from changing their answer. In light of this framework, Theorem 3.B shows that counterintuitive extreme types are the de facto "high types" in the knowledge screening setting.

5.3 Rewarding ignorance

In this subsection we fully characterize the conditions for the optimality of a two-threshold test, using a concavity-like but weaker condition. But first, we need to unavoidably define some additional notation.

Let $\widehat{v}_\omega$ denote the *concave envelope* of v_ω, i.e. the smallest concave function that lies above v_ω.

Note that a single threshold test need not be feasible for every given combination of NLR-threshold, undistorted state and prior. In other words, all values in $\bigcup_\omega v_\omega(\phi_\omega(T))$ need not be attainable for all π. As we shall see shortly, this is precisely what an optimal two-threshold test exploits – by effectively creating a lottery over a feasible and an *infeasible* but more lucrative single-threshold test.

Let $t_0^* := \max \arg\max_t V_0(\phi_0(t))$ and $t_1^* := \min \arg\max_t V_1(\phi_1(t))$. It can be easily verified using the expressions for V_1 above and analogous expressions for V_0, that t_1^* and t_0^* do not depend on π.

Now we are ready to state the main result of this subsection.

Proposition 3 (Characterization of conditions for TFU)**.** *The following are equivalent.*

- *There exists* $t \in (t_0^*, t_1^*)$ *such that* $\widehat{v}_1(\phi_1(t)) > V_1(\phi_1(t))$.
- *There exists* $t \in (t_0^*, t_1^*)$ *such that* $\widehat{v}_0(\phi_0(t)) > V_0(\phi_0(t))$.
- *If there exists a prior for which the optimal test is not "always pass" or "always fail", then there exists a prior for which a two-threshold test is optimal.*

Proposition 3 immediately leads to the following corollary.

Corollary 2. *If* V_1 *(equivalently,* V_0*) is concave, the optimal test for any prior has at most a single threshold.*

Note that Corollary 2 is implied by each of Proposition 1 and Proposition 3 but Proposition 3 does not imply Proposition 1.

Our interpretation of the above results is as follows. Proposition 3 essentially identifies a "convexity-like" property of the "effectiveness of screening" – as captured by the designer's value – in the *difficulty level* of the test, as necessary for the optimality of a TFU test. It says that in order for an *Unsure* option to be effective,

it is necessary that (within a range of weakly informative signals) marginally raising the bar of passing the test, when it is already quite high, improves the screening effectiveness *more*, than raising it when it is low. Raising the bar marginally entails excluding the "next" marginal signal from passing in a certain state, and therefore including it in the opposite state. At a high level, one way this can happen is when relative informativeness of the high type increases fast for extreme signals but not so much for moderate signals.

6 Universality of the simple True-False: Endogenous question selection

The comparative statics analysis of Section 5.2.1 leads to a natural next question - which priors maximize the principal's payoff? If the agent learns the same way about all facts, the prior captures the ex-ante uncertainty of each of them. Hence endogenizing the selection of the prior by the principal captures a natural feature of applications where the examiner or interviewer gets to select not only the evaluation scheme but also the questions or facts on which the candidate is to be evaluated.

Our main result in this section is that, given a choice over the prior, the principal always chooses "maximum uncertainty" – a *moderate* prior, in the sense of Section 5.2.1 – and consequently, offers a simple T-F test, regardless of other specifics.

Theorem 4 (Universality of the Simple True-False)**.** *Suppose the principal can choose the prior. The simple T-F test is optimal under some optimally chosen prior of the principal.*

Note that regularity or symmetry are not required for the above result to hold.

The main intuition for Theorem 4 is that an extreme prior gives too much away a priori. If the question is such that one of the two possible answers is too "obvious" ex-ante, the probability that examinees get it just by guessing is too much from the principal's perspective, which – recall – is an ex-ante one.

To understand why the above result holds, let us start from an extreme prior. Recall that the signal threshold does not depend on the prior, as long as the prior remains extreme. Hence for priors within the extreme range, the principal's payoff conditional on the counterintuitive state occurring remains invariant to the prior. As the prior grows moderate within this range, two forces are in action. First, the counterintuitive state grows more likely, which makes the principal better off. Second, the distortion in the obvious, distorted state reduces, which further aids the principal. This reduction happens because the belief at the constant signal threshold grows more moderate – closer to one half, where no distortion on the intensive margin is necessary. Due to these two forces, her expected gain from *both* states increases as the prior grows more moderate. In other words, the principal will never choose an

extreme prior because she always gains from making it more moderate. This leads to the first result above, which immediately leads to the second, by Theorem 2.A.

Some other notable features of the optimal prior and test under this setting are summarized in the result below.

Corollary 3. *Suppose, the principal can choose the prior. Then, in addition to Theorem 4, the following hold.*

- *The passing probability is at least one half.*
- *In addition, if the signal structures are symmetric and the principal's value function in each state is monotone, the principal optimal prior is* $\pi^* = \frac{1}{2}$.
- *If the problem is regular,* $\pi^* \in (\underline{\pi}, \overline{\pi})$ *where* $\underline{\pi}$ *and* $\overline{\pi}$ *are as defined in Theorem 2.A.*

7 Extensions

7.1 Agent observes quality

For our purposes, the setting where the agent does not know his quality, is not fundamentally different from the one, where he does. As we show in this section, all of our main results – Theorem 1 and slightly modified versions of theorems 2.A-2.B - hold even in this case.

First, even though the agent knows his quality, the principal obviously cannot screen by it alone, because in that case, in the absence of transfers or verification, there is no way to stop lower learning types from misreporting as higher ones. Hence she must screen by both quality and signal – combined into beliefs – just as in our main model. Hence basic characterizations of the set of incentive compatible and potentially optimal mechanisms in terms of beliefs – Theorem 1 – hold even in this case.

Proposition 4. *The optimal mechanism is a step function with at most two steps, regardless of whether the agent knows his quality.*

The essential elements of the second set of main results – theorems 2.A and 2.B – also hold true because the main forces are the same. The main difference between the two settings is that now a lower learning type needs a more extreme signal than a higher one in order to have an equally extreme belief. Hence each belief threshold of any optimal mechanism translates into *different* signal thresholds for the various learning types. Similarly as before, the "indifference at the threshold" principle must hold for each type, leading to a similar relationship between the threshold(s) and the extent of distortion(s). The additional condition is that each set of signal thresholds must give rise to the same belief.

To fix ideas, let us focus on the two learning type case. Let t_H and t_L be the signal-thresholds for high and low learning types in a single-threshold test. Using $\phi_e : [0,1] \to \mathbb{R}$ to denote the odds ratio of type $e \in \{H, L\}$ as a function of his signal, we must have:

$$\phi_H(t_H) = \phi_L(t_L) \iff \frac{f_{H1}(t_H)}{f_{H0}(t_H)} = \frac{f_{L1}(t_L)}{f_{L0}(t_L)}$$

As an illustration of the similarity between the two settings claimed above, let us compute the principal's value from a single threshold fail-if-incorrect test with a belief threshold below one half. With slight abuse of notation, let $t_H : [0,1] \to [0,1]$ be the implicit function capturing the high learning type's signal threshold as a function of that of the low one, i.e. $t_H : t \mapsto \phi_H^{-1}(\phi_L(t))$. Note that by strict increasingness of ϕ_H, t_H is well-defined. With that, the principal's value (scaled by r) from a single threshold fail-if-incorrect test with a belief threshold below one half with the low type's signal threshold t is given by:

$$\begin{aligned} V_-^F(t) = &\left(\pi \int_{t_H(t)}^{1} f_H(t'|1)dt' + (1-\pi)\left(\frac{\pi}{1-\pi}\right)\phi_H(t_H(t)) \int_{0}^{t_H(t)} f_H(t'|0)dt' \right) \\ &- \left(\pi \int_{t}^{1} f_H(t'|1)dt' + (1-\pi)\left(\frac{\pi}{1-\pi}\right)\phi_L(t) \int_{0}^{t} f_H(t'|0)dt' \right). \end{aligned} \tag{1}$$

From (1) it is clear that the characteristic features of the optimal test described in theorems 2.A and 2.B – such as invariance of the signal thresholds with the prior for a range of extreme priors and the distortion on the intensive margin reducing monotonically to zero as the prior goes from extreme to moderate – are preserved in this case as well. This is formalized below.

Proposition 5. *Suppose the agent observes his quality and the problem is regular. Then the following hold:*

- *There exist* $0 < \underline{\pi} \leq \overline{\pi} < 1$ *such that the type of the optimal test is always pass, never pass or given by the following:*

	$\pi < \underline{\pi}$	$\pi \in [\underline{\pi}, \overline{\pi}]$	$\pi > \overline{\pi}$
Cherry-picking, $V_0 < 0$	Fail-if-incorrect, penalize False	Simple T-F	Fail-if-incorrect, penalize True
Lemon-dropping, $V_0 \geq 0$	Pass-if-correct, Bonus for True	Simple T-F	Pass-if-correct, Bonus for False

Table 5: Optimal tests across markets and priors

Moreover, if the signal structures are symmetric, $\underline{\pi} + \overline{\pi} = 1$ and $\underline{\pi} < \frac{1}{2}$.

- *When the prior is extreme, as it grows more moderate, the signal threshold of the optimal test for each of the learning types remains constant, its (common) belief threshold grows more moderate and the (common) distortion in the allocation in the a priori obvious state reduces.*

- *When the prior is moderate, as it increases, the signal threshold in the optimal test for each of the learning types increases and its belief threshold remains constant at one half. Consequently there is no distortion in the allocation in either state.*

7.2 Alternative timings of the game

7.2.1 The informed principal game

In the main model we assume the principal does not know the answer to the question before choosing the test. However, in many real world scenarios such as teachers setting exam questions, this need not be the case. When the principal knows the true state before choosing the test, we have an *informed principal problem* (Myerson (1983),Maskin and Tirole (1990)). In this section we show that under regularity, the ex-ante optimal mechanism is an equilibrium of the informed principal game even in that case. Moreover, under mild additional assumptions, the main qualitative properties of the optimal tests highlighted throughout this paper hold for optimal tests arising in *any* other equilibrium of the informed principal game.

The solution concept we use is that of *undominated mechanisms*, as introduced by Myerson (1983). In our setting, a mechanism $(\widehat{a}_1, \widehat{a}_0)$ is *undominated* if and only if there is no other mechanism under which the principal's payoff in *each* state is weakly greater than under $(\widehat{a}_1, \widehat{a}_0)$ and strictly greater than $(\widehat{a}_1, \widehat{a}_0)$ in at least one state. This immediately leads to the following observation.

Fact 2. *A mechanism $a^0, a^1 : P \to [0,1]$ is an undominated mechanism if only if it solves the following for some $\pi_P \in [0,1]$.*

$$\max_{a_1, a_0 \in [0,1]^P} \quad \pi_P \int_p v_1(p) a_1(p) dF(p|1) + (1 - \pi_P) \int_p v_0(p) a_0(p) dF(p|0) \qquad (1)$$

s.t. Feas and IC.

Fact 2 combined with Theorem 1 immediately tells us that any undominated mechanism must have the same at-most-two-thresholds structure established earlier. Therefore, analogously as Theorem 1, we observe the following:

Proposition 6. *Any undominated mechanism (a_1, a_0) consists of step functions with at most two steps, where the a_1 and a_0 change at the same belief(s).*

As such, we reference undominated mechanisms by the relative weight they put on the principal's value when the realized state is 1. Specifically, an undominated mechanism $\widetilde{\pi}$ is one which maximizes (1) with $\pi_P = \widetilde{\pi}$. Clearly, a mechanism is ex-ante optimal if and only if it solves (1) for $\pi_P = \pi$.

Next we show that under an (appropriately strengthened) regularity condition, analogs of theorems 2.A and 2.B hold for *any* undominated mechanism in the informed principal game as well.

"Regularity" in the informed principal game

In the main analysis we called a problem *regular* if the corresponding virtual value was increasing whenever $\pi_P = \pi_A$. Analogously, we call a problem *strongly regular* if the corresponding virtual value is increasing for *all* (π_P, π_A) pairs.

Clearly, strong regularity implies regularity.

An example of a class of natural settings satisfying strong regularity is given below.

Observation 1. *Consider Special Case 1. Suppose further that the high quality type's signal structure is linear – $f(t) = 1 - b + 2bt, b \in [0, 1]$ – then the problem is strongly regular whenever it is non-trivial for any $\gamma \geq 60\%$, regardless of other parameters.*

The theorem below shows that for a symmetric, strongly regular problem, our main characterization, Theorem 2.A still holds, as long as the principal's losses from the low quality types are "not too high". In the following theorem we use *loss* to mean the absolute value of the principal's payoff from agents whose quality is negative.

Theorem 5. *If the problem is strongly regular and* $\max_{t\in[0,1]} \min\{m_1(t), m_0(t)\} \geq 0$, *then for any undominated mechanism π_P, the type of the optimal test is always pass, never pass or given by the following:*

	$\pi_A < \underline{\pi}(\pi_P)$	$\pi_A \in [\underline{\pi}(\pi_P), \overline{\pi}(\pi_P)]$	$\pi_A > \overline{\pi}(\pi_P)$
Cherry-picking, $u_0 < 0$	Fail-if-incorrect, penalize False	Simple T-F	Fail-if-incorrect, penalize True
Lemon-dropping, $u_0 > 0$	Pass-if-correct, bonus for True	Simple T-F	Pass-if-correct, bonus for False

Table 6: Optimal tests across markets, priors and equilibria of the informed principal game

In particular, for any problem $\mathcal{P}$ there exists $k \in [0, 1)$ such that if all losses are

reduced by the same proportion k, the condition $\max_{t\in[0,1]} \min\{m_1(t), m_0(t)\} \geq 0$ *is satisfied.*

Other equilibrium concepts

It may be worth mentioning that there are solution concepts other than undominated mechanisms as well, under which the ex-ante optimal mechanism we characterized remains an equilibrium mechanism of the informed principal game. In particular, we consider the concept of *core mechanisms*, as defined by Myerson (1983). In our context, a core mechanism is one for which neither of the two principal "types" can do better by deviating and revealing her type.

Proposition 7. *Suppose the problem is regular. Then the ex-ante optimal mechanism is a core mechanism.*

The main ideas behind the above result are as follows. Under regularity the optimal tests have a single threshold. The common threshold is chosen *trading off* principal's expected values in the two states. But each state nevertheless "skims the cream" – fails (or fails with a greater probability, in the lemon-dropping case) the signals at its *respective worse extreme*. Thus, under such optimal tests the principal strictly benefits from discrimination in *each* state. On the other hand if the principal reveals the state, no discrimination is possible. Therefore the principal of each "type" must prefer the ex-ante optimal mechanism over revealing her type, i.e. the state itself. The details are in the Appendix.

7.2.2 Strategic learning by the principal

In this section we consider two alternative timings where the principal makes a strategic choice at the beginning of the game regarding whether or not to learn the state. We show that in both cases she chooses the latter. We also show, by way of an example, that when she is allowed to choose from the full set of possible experiments about the state however, her optimal choice can be intermediate.

The first alternative timing is as follows.

1. The principal chooses between acquiring a fully informative and a fully uninformative signal about the state.

2. The agent observes her choice.

3. The principal's signal is realized, unobserved by the agent.

4. If she has chosen the informative test in Step 1, the game given in Section 7.2.1 ensues. If she has chosen the uninformative test instead, that of Section 2 ensues.

In the second alternative timing, the agent also observes the principal's signal.

1-3. As given above.

4. The agent observes the principal's signal.

4. The game given in Section 2 ensues, but now with the potentially updated fully revealing common prior, $\pi \in \{0, 1\}$.

We show that in either of these two cases, the principal chooses not to learn.

Proposition 8. *For any of the two alternative timings given above, the principal chooses the uninformative signal in Step 1.*

The reasoning is as follows. In the second case, if the principal chooses to learn the state, the agent also learns it. Clearly, in this case there can be no screening thereafter, which the principal cannot prefer. Now suppose the agent does not observe her signal. In this case, if the principal has chosen to learn the state in the first stage, an informed principal game ensues, as in Section 7.2.1. As we showed there, all equilibria of that game are weakly worse for the principal from her ex-ante perspective, than if she can commit to the mechanism without learning the state, and the agent knows that she has not learnt the state.

7.3 Efficient tests and questions

In this subsection we show that the class of optimal tests characterized in Theorem 1 remains optimal even when the test-designer cares – not only about the screening effectiveness - but also about the test-taker's payoffs.

Let us assume the test-designer wants to maximize a convex combination of her own and the agent's payoff – akin to a "social welfare" function. Let us call the class of tests which maximize such a welfare function, "efficient" tests. Relooking at her original program 1, and considering our main analysis, it is clear that in this case the class of optimal tests remains the same as that characterized in Theorem 1. This is because even in this case, the welfare function remains linear in the mechanism, (a_1, a_0), and therefore the set of potential optimal mechanisms remains the set of extreme points of the feasible set, which does not change. This is formalized below.

Proposition 9 (Analog of Theorem 1)**.** *Any efficient test* (a_1, a_0) *consists of step functions with at most two steps, where the* a_1 *and* a_0 *change at the same belief(s).*

In our microfounded model from Section 5, the principal caring about social welfare is equivalent to her value from each learning type going up. Therefore all of our results in sections 5 and 6 remain qualitatively valid too.

7.4 Multi-question tests

In this subsection we consider the design of a test consisting of $n \in \mathbb{N}$ questions, each about a binary fact. We show that under a natural design restriction, our results apply to the design of such tests as well.

Now we have N binary states – let us call them states 1 to N – and 2^N possible state realizations. Let $\omega := (\omega_1, \cdots, \omega_N)$, where $\omega_i \in \{0,1\}$ for each i, denote the typical state realization. $\therefore \Omega = \{\omega : \omega = (\omega_1, \cdots, \omega_N), \omega_i \in \{0,1\} \ \forall \ i\}$.

Our belief space is the $(2^N - 1)$-simplex, $P := \{p \in [0,1]^{2^N} : \sum_{\omega \in \Omega} p_\omega = 1\}$. Let p^i denote the agent's marginal belief on the i-th state, i.e. $p^i = \sum_{\omega_{-i} \in [0,1]^{N-1}} p_{(\omega_{-i}, \omega_i)}$, where $\omega_i = 1$.

The analog of the mechanism (a_1, a_0) in this context would be $\mathbf{a} : P \to [0,1]^{2^N}$. a family of reward functions, one for each state. We can interpret $\mathbf{a}$ as passing probabilities, as in the main model, but for the rest of the section we would call it *scores*, which is more interpretable in this context.

We impose the design restriction that the total score has to be a weighted average of scores for each question and that each question must be scored based on the agent's knowledge of that question only, i.e., his marginal belief over the binary state representing that question. Formally:

Restriction 1. *There exist functions* $\mathbf{a}^i_{\omega_i} : [0,1] \to [0,1]$ *and weights* $s_i \in [0,1] \ \forall \ i \in \{1, \cdots, N\}, \sum_i s_i = 1$, *such that for all* $p \in P, \omega \in \Omega$,

$$\mathbf{a}_\omega(p) = \sum_{i=1}^{N} s_i \mathbf{a}^i_{\omega_i}(p^i)$$

To reiterate, we maintain the assumption, that the agent cares about the total score only. Using $U(p; p')$ to denote the indirect utility of an agent with belief p, from reporting p', formally this means:[26]

$$U(p; p') = \sum_{\omega} p_\omega \mathbf{a}_\omega(p')$$

The principal, as before, wants to maximize her ex-ante payoff, given her interim payoff function which depends on the belief-state combination.

Our main result in this subsection is that, when the principal is restricted to design essentially a separate test, $\{(a^i_1, a^i_0)\}$, for each question i, each of these tests take the same format of making the agent pick from at most three options, as in our main characterization. The analog of Thorem 1, therefore, would be:

[26] Note that this assumption is equivalent to imposing risk neutrality over scores, if we interpret the reward as some numerical score instead of passing probabilities.

Proposition 10. *Under any optimal mechanism* $\{(a_1^i, a_0^i)\}_{i=1}^N$, a_1^i *and* a_0^i *consist of step functions with at most two steps, where the* a_1^i *and* a_0^i *change at the same belief(s), for all* $i \in \{1, \cdots, N\}$.

The key to the above result is that under the design restriction 1, incentive compatibility is equivalent to question-wise incentive compatibility – not wanting to misreport one's marginal belief over each of the states. A corollary of design restriction 1 is also that under it, the mechanism cannot distinguish among agents with the same marginal beliefs for each state.

7.5 Risk averse agents

In this paper we have maintained that the passing probabilities can be interpreted as number of points, giving a natural interpretation of intermediate passing probabilities as partial credit. However, this claim rests on the assumption that the agent is an expected points maximizer, i.e., risk neutral.

Suppose, instead, that the agent is risk averse but the principal is still risk neutral, i.e. her expected payoff is linear in the rewards of the agent (a_1 and a_0). In this case our main claim of simplicity – implementability of the optimal tests with at most three options – no longer holds in general. The reasoning is as follows.

In this case the appropriate primitive to work with is the agent's utility from a given number of points, from which the number of points can be backed out, to be plugged into the principal's payoff. Straightforward algebra shows that in this case, her expected payoff is no longer necessarily convex in the aforementioned utility function of the agent. Therefore optimizers are no longer necessarily extremal.

The main takeaway from the above insight is that when testing risk averse agents, it may be optimal to give them more options – upto infinitely many – to express their beliefs about the answer to a question. For future work, it may be interesting to see how heterogeneous and/or private risk aversion affects the optimal test – an analysis which has implications to the debate on the differential impacts of negative marking on men and women test-takers, who systematically differ in terms of their risk aversion Baldiga (2014); Saygin and Atwater (2021); Coffman and Klinowski (2020).

7.6 Forcing tests to be "fair"

One of the potentially surprising results to come out of our main analysis is that optimal tests could be "unfair", in the sense that they could pass incorrect answers or fail correct answers (Section 5.2.1). However, in the real world, it is hard to justify failing a test-taker who gives a correct answer, even if such a grading scheme offers the optimally effective screening of the relevant underlying quality of the candidate that the test-designer cares about. The main result of this subsection

shows that such a restriction – that "fully correct" answers must be passed with a 100 % probability – can be accommodated in our framework, and that they keep the multiple-choice structure of the test unaltered, give rise to potentially just one additional option.

We model the "fairness" restriction in the following way. We assume that the "most correct" belief must be passed regardless of the correct answer. In terms of our formal mechanism (a_1, a_0) introduced in our main analysis in Section 4.2, this imposes the additional restrictions:

$$a_0(\underline{p}) = 1,$$
$$a_1(\overline{p}) = 1.$$

We say a test (a_1, a_0) is *fair*[27] if the above restrictions hold.

So the principal's problem, (1), becomes:

$$\max_{a_1, a_0 \in [0,1]^P} \quad \pi \int_p v_1(p) a_1(p) dF(p|1) + (1-\pi) \int_p v_0(p) a_0(p) dF(p|0) \tag{1}$$
$$\text{s.t. } a_1, a_0 \in [0,1]. \tag{Feas}$$
$$p a_1(p) + (1-p) a_0(p) \geq p a_1(p') + (1-p) a_0(p'), \forall p, p' \in P \tag{IC}$$
$$a_0(\underline{p}) = 1, \tag{Pass correct 0}$$
$$a_1(\overline{p}) = 1. \tag{Pass correct 1}$$

It turns out, even With this new restriction, the class of optimal tests remains the same, as captured by the following result.

Proposition 11 (Analog of Theorem 1)**.** *The optimal fair mechanism* (a_1, a_0)*, which solves* (1)*, consists of step functions with at most two steps, where the* a_1 *and* a_0 *change at the same belief(s).*

The proof is analogous to that of Theorem 1.

8 Conclusion

In this paper we analyze the problem of a principal trying to maximize the probability of accepting a good type of agent and minimize that of accepting a bad type of agent, by designing a test of the agent's knowledge of an unknown state,

[27] Note that this nomenclature is unrelated to the equity-based notion of fairness often used in studying multi-agent mechanisms, where fairness captures some notion of "equal treatment of equals".

which is correlated with his quality. Interviews and other knowledge-based tests frequently employed in passing settings constitute our main application. We show that the optimal tests take a simple "pick the correct answer" form, observed often in reality. In general there can be at most three options – which can be thought of as True/False/Uncertain. Under some natural regularity conditions, the optimal test features at most two options – True and False. However, we show that due to agency issues, giving a correct answer need not always earn the same credit. The partial credits earned by the answer depends on the correct answer itself. This leads to over-rewarding of "counterintuitive" – or a priori unlikely – answers and under-rewarding of "obvious" (a priori likely) ones, compared to the efficient (full information) test. When the principal can choose the question – modeled as choosing the prior for the unknown state on which the agent is quizzed – she naturally prefers to avoid these distortions. Therefore in that case, we show, she chooses the prior in a way such that there is no distortion at either extreme, even without regularity conditions. Under regularity conditions, when the principal can choose the prior, the optimal test always takes the form of a simple, unweighted True-False question.

References

J. D. Abernethy and R. M. Frongillo. A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pages 27–1. JMLR Workshop and Conference Proceedings, 2012.

Ş. P. Akyol, J. Key, and K. Krishna. Hit or miss? test taking behavior in multiple choice exams. Technical report, National Bureau of Economic Research, 2016.

K. Baldiga. Gender differences in willingness to guess. *Management Science*, 60(2): 434–448, 2014.

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

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

C. P. Chambers and N. S. Lambert. Dynamic belief elicitation. *Econometrica*, 89 (1):375–414, 2021.

K. B. Coffman and D. Klinowski. The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*, 117 (16):8794–8803, 2020.

R. Deb, M. M. Pai, and M. Said. Evaluating strategic forecasters. *American Economic Review*, 108(10):3057–3103, 2018.

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

J. Glazer and A. Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6): 1715–1736, 2004.

N. Hancart. Designing the optimal menu of tests. 2022.

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

R. L. Karandikar. On multiple choice tests and negative marking. *Current Science*, 99(8):1042–1045, 2010.

N. S. Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 2011.

E. Lesage, M. Valcke, and E. Sabbe. Scoring methods for multiple choice assessment in higher education–is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, 39(3):188–193, 2013.

Y. Li, J. D. Hartline, L. Shan, and Y. Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 988–989, 2022.

I. Marinovic, M. Ottaviani, and P. Sorensen. Forecasters' objectives and strategies. In *Handbook of economic forecasting*, volume 2, pages 690–720. Elsevier, 2013.

E. Maskin and J. Tirole. The principal-agent relationship with an informed principal: The case of private values. *Econometrica: Journal of the Econometric Society*, pages 379–409, 1990.

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

M. Mussa and S. Rosen. Monopoly and product quality. *Journal of Economic theory*, 18(2):301–317, 1978.

R. B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1): 58–73, 1981.

R. B. Myerson. Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, pages 1767–1797, 1983.

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

M. Ottaviani and P. N. Sørensen. Professional advice. *Journal of Economic Theory*, 126(1):120–142, 2006.

J.-C. Rochet and P. Choné. Ironing, sweeping, and multidimensional screening. *Econometrica*, pages 783–826, 1998.

J.-C. Rochet and L. A. Stole. Nonlinear pricing with random participation. *The Review of Economic Studies*, 69(1):277–311, 2002.

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

P. O. Saygin and A. Atwater. Gender differences in leaving questions blank on high-stakes standardized tests. *Economics of Education Review*, 84:102162, 2021.

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

G. Winkler. Extreme points of moment sets. *Mathematics of Operations Research*, 13(4):581–587, 1988.

A Example: Additional calculations

In general, for any given prior π, the mapping between beliefs and signals is given by:

$$\mu(t) = \frac{\pi(2t+1)}{\pi(2t+1) + (1-\pi)(2(1-t)+1)} \tag{1}$$

This gives us the range of beliefs, $P = \left[\frac{\pi}{3-2\pi}, \frac{3\pi}{2\pi+1}\right]$.

Using (1) and simplifying, we have:

$$\mu^{-1}(p) = \frac{2}{1 + \left(\frac{\pi}{1-\pi}\right)\left(\frac{1-p}{p}\right)} - \frac{1}{2} \tag{2}$$

Differentiating,

$$\mu^{-1\prime}(p) = \frac{2}{\left(1 + \left(\frac{\pi}{1-\pi}\right)(1-p)\right)^2} \tag{3}$$

Also, using (1), in general, the teacher's value from passing belief p in the two states:

$$V_1(p) = \frac{u_H \times \frac{1}{2} \times f_{1H}(\mu^{-1}(p))dt + u_L \times \frac{1}{2} \times f_{1L}(\mu^{-1}(p))dt}{f(p|1)dp}$$
$$V_0(p) = \frac{u_H \times \frac{1}{2} \times f_{0H}(\mu^{-1}(p))dt + u_L \times \frac{1}{2} \times f_{0L}(\mu^{-1}(p))dt}{f(p|0)dp}$$

The above gives us the "weighted" value of belief p in state 1:

$$\begin{aligned} V_1(p)f(p|1)dp &= \left(f_{1H}(\mu^{-1}(p)) - \frac{1}{2}\right)dt \\ &= \left(2\mu^{-1}(p) - \frac{1}{2}\right)\mu^{-1\prime}(p)dp \\ &= \frac{\left(\frac{8}{1+\left(\frac{\pi}{1-\pi}\right)\left(\frac{1-p}{p}\right)} - 3\right)}{\left(p + \left(\frac{\pi}{1-\pi}\right)(1-p)\right)^2}dp \end{aligned}$$

$$\begin{aligned} V_0(p)f(p|0)dp &= \left(f_{0H}(\mu^{-1}(p)) - \frac{1}{2}\right)dt \\ &= \left(2(1-\mu^{-1}(p)) - \frac{1}{2}\right)\mu^{-1\prime}(p)dp \\ &= \frac{\left(5 - \frac{8}{1+\left(\frac{\pi}{1-\pi}\right)\left(\frac{1-p}{p}\right)}\right)}{\left(p + \left(\frac{\pi}{1-\pi}\right)(1-p)\right)^2}dp \end{aligned}$$

B Formula for the optimal test

We start by specifying expressions for the principal's value from the various types of tests, which would be used in deriving the additional details of its structure which we provide in this section.

Let $V_-^i : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ and $V_+^i : \left[\frac{1}{2}, 1\right] \to \mathbb{R}$ denote the principal's value from a single-threshold test of type $i \in \{P, F\}$ - where P and F stand for pass-if-correct and fail-if-incorrect test types respectively - as a function of its (belief) threshold. Clearly,

$$\begin{aligned} V_-^P(p_0) &= \pi \int_{p_0}^{\overline{p}} v^1(p)dp + (1-\pi)\left[\int_{\underline{p}}^{p_0} v^0(p)dp + \left(1 - \frac{p_0}{1-p_0}\right)\int_{p_0}^{\overline{p}} v^0(p)dp\right], \\ V_-^F(p_0) &= \pi \int_{p_0}^{\overline{p}} v^1(p)dp + (1-\pi)\left(\frac{p_0}{1-p_0}\right)\int_{\underline{p}}^{p_0} v^0(p)dp, \\ V_+^F(p_0) &= \pi\left[\left(1 - \frac{1-p_0}{p_0}\right)\int_{\underline{p}}^{p_0} v^1(p)dp + \int_{p_0}^{\overline{p}} v^1(p)dp\right] + (1-\pi)\int_{\underline{p}}^{p_0} v^0(p)dp, \\ V_-^F(p_0) &= \pi\left(\frac{1-p_0}{p_0}\right)\int_{p_0}^{\overline{p}} v^1(p)dp + (1-\pi)\int_{\underline{p}}^{p_0} v^0(p)dp. \end{aligned} \tag{Tests}$$

B.1 Dual threshold tests: Ironing

Sometimes, if the principal's favorite common threshold bang-bang mechanism has a threshold that is too extreme - requiring a large adjustment either on the intensive or on the extensive margins - it may be better for her to introduce an *additional* threshold, instead of making these adjustments while maintaining a single threshold (See Figure 1b). Mathematically, the optimal mechanism features two thresholds when the optimal knowledge premium (q) is a convex combination of the knowledge premia of two (non-incentive compatible) bang-bang mechanisms. The *average* of the thresholds for these two bang-bang mechanisms is equal to one half. In other words, incentive compatibility conditions require that the optimal value is the *concavified* value of bang-bang mechanisms as a function of their thresholds, at $\frac{1}{2}$.

In order to simplify the analysis of the two-threshold case, we redefine the virtual value by picking $p_0 = 0$ in (4). As discussed earlier, the principal's value from any incentive compatible mechanism is given by (4) for *any* choice of p_0. If we choose $p_0 = 0$, the second term in (4) vanishes, leaving us with only one virtual value, $\chi := \chi_1$, i.e. our virtual value is given by:

$$\chi(p) = \pi(1-p)v^1(p) - (1-\pi)pv^0(p) + \int_{p' \geq p} v(p')dp' \tag{VV}$$

The value of the principal from a bang-bang mechanism with threshold p is clearly $\mathbb{E}v + \int_p^{\overline{p}} \chi(p')dp' - \int_{\underline{p}}^{p} \chi(p')dp'$. Hence, per the above discussion, her maximized value from a dual threshold test is $\mathbb{E}v + \int_p^{\overline{p}} \widetilde{\chi}(p')dp' - \int_{\underline{p}}^{p} \widetilde{\chi}(p')dp'$, where $\widetilde{\chi}$ is the ironed virtual value, where ironing is used in the sense of Myerson (1981).

With that background we can describe the optimal mechanism more specifically, as given below.

[General solution] Either the optimal mechanism is constant with respect to both the state and the agent's report, or the principal's value is given by the solution to the following one-dimensional maximization problem:

$$\max_{p_0} V(p_0)$$

where

$$V(p_0) = \begin{cases} \max\{V_-^F(p_0), V_-^P(p_0)\}, \text{ for } p_0 < 1/2 \\ \mathbb{E}v + \int_{p_0}^{\overline{p}} \widetilde{\chi}(p')dp' - \int_{\underline{p}}^{p_0} \widetilde{\chi}(p')dp', \text{ for } p_0 = 1/2 \\ \max\{V_+^F(p_0), V_+^P(p_0)\}, \text{ for } p_0 > 1/2 \end{cases} \tag{1}$$

Note that for single threshold tests, the threshold type is indeed p_0 - the type introduced earlier as the "lowest" type - as hinted by the above notation.

C Discussion

In this section we provide interpretations of the various objects in our framing of the screening problem – such as the object we interpret as analogous to "allocation" in the standard screening problem, and virtual value.

Note that (1) can be written alternatively as follows, fixing any $p_0 \in P$ as the "base type" – the optimal q function will not depend on it.

$$U(p) = U(p_0) + \int_{p_0}^{p} q\left(p'\right) dp', \ \forall\, p, p_0 \in P. \tag{1}$$

For *solving* the problem, it is simplest to set p_0 to be equal to one of the extremes of the type space – $\overline{p}$ or $\underline{p}$ – as we did in the main text. The reason we still introduce the above more general formulation in this section, is that it is helpful in interpreting the various objects in our model.

Using the fact that $U(p) = pa_1(p) + (1-p)a_0(p)$ in conjunction with Lemma 1 gives us the following version of (Integral Formulas):

$$\begin{aligned} a_0(p) &= \left(U(p_0) + \int_{p_0}^{p} q\left(p'\right) dp'\right) - pq(p) \\ a_1(p) &= \left(U(p_0) + \int_{p_0}^{p} q\left(p'\right) dp'\right) + (1-p)q(p) \end{aligned} \qquad \text{(Generalized integral Formulas)}$$

C.1 Significance of the model elements

C.1.1 The base type

As seen in the previous section, we can pick any base type p_0 without affecting the optimal mechanism. Let us interpret p_0 as (one of) the "lowest" type(s) - that is, (one of) the type(s) with the lowest interim passing rate under the optimal mechanism. Hence by definition $\int_{p_0}^{p} q\left(p'\right) dp' \gtreqless 0$ according to $p \gtreqless p_0$. This, combined with the monotonicity of q, must mean $q(p) \gtrless 0 \iff p \gtrless p_0$. Hence we can formally define p_0 as:

$$p_0 := \sup\{p : q(p) \leq 0\} \tag{1}$$

That is, the p_0 is the (necessarily unique) type where the passing rate changes from being weakly greater in state 0 to being strictly greater in state 1.[28] Hence,

[28] Note that all types p with $q(p) = 0$ are "lowest" - they all have the same lowest interim passing rate.

the mechanism is treating types $p < p_0$ as relatively more suited to be passd when the correct answer is 0, because they are relatively certain that the state is 0 - let us call them the "left types" - and the reverse for $p > p_0$ - who we would call the "right types". This is consistent with p_0 being the "lowest" type - who the mechanism treats as "ignorant".

With this interpretation of the base type in mind, we turn to interpreting the components of $\widehat{a}_1$ and $\widehat{a}_0$, as given by (Integral Formulas). The first component in either of the expressions is the requisite interim rate at which each type must be passd, as dictated by incentive constraints (equation 1), which is the same in both the states. In addition, each rate is distorted in a direction and by an amount consistent with the type. Specifically, each type - left or right - is passd at a rate higher (lower) than its interim passing rate when the correct answer is the one (is not the one) towards which he is biased. The amounts by which the interim passing rate is adjusted up or down in case of a correct or wrong answer respectively, grows more extreme in proportion to the extremity of the belief. In light of the above framing, we call the absolute value of $q(p)$ type p's *knowledge premium* - the premium in passing probabilities the type receives for their bias towards the correct answer, compared to the wrong answer.

C.1.2 The virtual value

Using (Generalized integral Formulas), we can express the principal's ex-ante value in terms of her "virtual value" functions – χ_0 and χ_1 for beliefs more extreme than p_0 towards states 0 and 1 respectively – as follows:

$$\mathbb{V}(a_1, a_0) = U(p_0)\mathbb{E}v + \int_{p \leq p_0} \chi_0(p)q(p)dp + \int_{p \geq p_0} \chi_1(p)q(p)dp \tag{2}$$

We provide the derivation of (4) and explicit formulas for χ_0 and χ_1 in Section D.1.3.

Unlike the standard monopoly problem, the goals of our principal and agent (recall that they represent an principal and a job seeker, in our leading application) are not strictly adversarial - taking surplus away from the agent is not the only way the principal can enrich herself. In the standard monopoly case, no type can earn *less* under the incentive-constrained solution, than their full information surplus, which is zero. That is not the case here. Hence, as we will show going forward, information rent is not always *paid* by the principal, but sometimes *earned*.

In our model, the virtual value of a type is the value to the principal per unit of knowledge premium allocated to that type. χ_0 and χ_1 - as in (3) - capture the virtual values of the left and right types respectively. The virtual values from the left and right types can be decomposed into "direct" and information rent effects as follows:

$$\begin{aligned}\chi_0(p) &= \underbrace{\pi(1-p)v^1(p) - (1-\pi)pv^0(p)}_{\text{Direct effect}} + \underbrace{\int_{p' \leq p} v(p')dp'}_{\text{Information rent}} \\ \chi_1(p) &= \underbrace{\pi(1-p)v^1(p) - (1-\pi)pv^0(p)}_{\text{Direct effect}} + \underbrace{\int_{p' \geq p} v(p')dp'}_{\text{Information rent}}\end{aligned} \tag{3}$$

In order to understand its direct and information rent components, let us decompose the principal's value per unit of incremental change in it as follows:

$$\begin{aligned}&\pi v^1(p)a_1(p) + (1-\pi)v^0(p)a_0(p) \\ &= \underbrace{v(p)U(p)}_{\text{Source of information rent}} + \underbrace{(\pi(1-p)v^1(p) - (1-\pi)pv^0(p))}_{\text{Direct effect}} q(p), \\ &\text{putting } a_1(p) = a_0(p) + q(p) \text{ and } a_0(p) = U(p) - pq(p)\end{aligned}$$

The first term above is the interim unconditional (on the state) value from each type times the interim passing rate of that type - it is the expected value the principal would get from type p if she had no information about the correct answer. The second part then, is the "premium" she gets due to knowing the right answer, which the agent does not know - the value-weighted knowledge premium. Note that this premium can never be negative under the unconstrained solution, though it can be under the constrained solution.

From the above decomposition it is clear that the direct value of type p to the principal should be the per unit value to her of increasing $q(p)$ by a small amount while *keeping $U(p)$ unchanged*. If we want to increase $q(p)$ by ϵ while keeping type p indifferent between knowledge premia $q(p)$ and $q(p) + \epsilon$, then we need to implement this change by changing $a_1(p)$ and $a_0(p)$ in the right proportions. In particular, we need to increase $a_1(p)$ by $\epsilon(1-p)$ while decreasing $a_0(p)$ by $p\epsilon$. The net effect of this adjustment on the interim passing rate of type p is $p \times (1-p)\epsilon - (1-p) \times p\epsilon = 0$. The value of this change to the principal is clearly $\pi \times v^1(p) \times$ change in $a_1(p) + (1-\pi) \times v^0(p) \times$ change in $a_0(p) = (\pi(1-p)v^1(p) - (1-\pi)pv^0(p))\epsilon$, giving us the effect per unit of incremental change in $q(p)$ as $(\pi(1-p)v^1(p) - (1-\pi)pv^0(p))$, as we see from the decomposition above.

The information rent part can be further decomposed as:

$$v(p)U(p) = \underbrace{v(p)U(p_0)}_{\text{Info rent: Constant part}} + \underbrace{v(p)\int_{p_0}^{p} q(p')dp'}_{\text{Info rent: Variable part}}$$

The first part, when averaged over all types, gives rise to the first term in (4). It is easy to see why - under the constrained solution, every type must be passd at an interim rate of *at least* $U(p_0)$, regardless of the state, which gives the principal the baseline value of $\mathbb{E}v \times U(p_0)$, before she utilizes any of her knowledge of the state via q. The second, variable (with chosen q) part of information rent gives rise to the integral terms in the expressions for χ_0 and χ_1, as in (3). As we can see, these terms result from the impact of increasing the knowledge premium of type p by ϵ, on more extreme types. Naturally, incentives demand that if type p is rewarded with an increase in its knowledge premium of ϵ, so must be types more extreme than it, at the least. Mathematically, this is the familiar monotonicity condition on q. By (1), this means types *more extreme* than it must passd at the interim rate of ϵ more. The unconditional interim value from any more extreme type p' is $v(p')$, which is accrued to the principal due to this change, as captured by the two integral terms in (3).

C.1.3 Trade-off between guaranteed passing rate and knowledge premia

Note that both the components $\int\limits_{p\leq p_0} \chi_0(p)q(p)dp$ and $\int\limits_{p\geq p_0} \chi_1(p)q(p)dp$ in (4) must be positive at the principal's optimal mechanism, because if not, the principal can do better by setting $a_1(p) = a_0(p) = U(p_0)$ for all $p < p_0$ or $p \geq p_0$, depending on whether $\int\limits_{p\leq p_0} \chi_0(p)q(p)dp$ or $\int\limits_{p\geq p_0} \chi_1(p)q(p)dp$ is negative, respectively. Hence, the principal always wants to make the knowledge premium more extreme, if possible. This leads to a trade-off between the guaranteed interim passing rate the principal offers - $U(p_0)$ - and the knowledge premium, $|q|$. To see why, consider the following cases.

Suppose $\mathbb{E}v > 0$. So the principal wants to give a high guaranteed passing rate - driven by the first term in (4). However, if $U(p_0)$ is too high, she misses out on more accurate passing because she is left with little room to incentivize greater information revelation by extreme types (due to (1)). In other words, she loses out on the full discerning power of the knowledge premium because she cannot make $\widehat{a}_1$ for right types and $\widehat{a}_0$ for left types as high as she would like.

Similarly if $\mathbb{E}v < 0$ the exact opposite happens - in this case if $U(p_0)$ is too low, the principal cannot make $\widehat{a}_0$ for right types and $\widehat{a}_1$ for left types as low as she would like.

C.2 Regularity

In this subsection we provide some examples of natural classes of settings where our regularity assumption holds.

We say the problem is *trivial*, if the principal's value from every agent type in every state is non-positive, i.e. if $m_\omega \leq 0$ – equivalently $v_\omega \leq 0$ – for $\omega \in \{0, 1\}$.

Clearly, this case is trivial because in this case the principal's optimized value is zero - she passes no one in any state because every type earns her non-positive value. Otherwise we call the problem *non-trivial*.

Special Case 1. *Suppose the learning types binary and the signal structure of the high quality agent is symmetric, i.e.* $f_{H1}(t) = f_H(1-t|0) = f(t)$ *for some increasing* $f : [0,1] \to \mathbb{R}_+$. *Suppose further, that that of the low quality agent takes the following special form: in each state, the low quality agent gets the same signal as the high quality agent with some probability,* $\gamma \in [0,1)$, *and receives the uninformative signal* $f_\emptyset : [0,1] \to \mathbb{R}_+, f_\emptyset(t) = 1 \ \forall \ t$, *otherwise. That is, the low types signal structure is given by:* $f_{L\omega}(t) = \gamma f_{H\omega}(t) + 1 - \gamma$, $\omega \in \{0,1\}$. *Without loss we normalize the principal's payoff from the high learning type to 1 and assume that from the low learning type to be* $-u, u > 0$.

The following proposition provides natural classes of examples within Special Case 1 which satisfy regularity.

Proposition 12. *Consider Special Case 1.*

- *As long as the difference* $(f(t) - f(1-t))$ *is not too convex whenever it is positive - in particular, as long as the elasticity of the rate of change in* $(f(t) - f(1-t))$ *w.r.t.* $(f(t) - f(1-t))$ *is weakly lower than 1 whenever* $(f(t) - f(1-t))$ *is positive - for low enough* r, *the problem is regular whenever it is non-trivial, regardless of other parameters.*

- *As an example, if the low quality type is uninformed and the high quality type's signal structure is linear -* $f(t) = 1 - b + 2bt, b \in [0,1]$ *- then the problem is regular for any* $r \leq \frac{1}{2}$, $b \in [0,1]$ *and* $u > 0$.

- *When the signal structure follows a power law distribution and the low type is uninformed -* $f(t) = (m+1)t^m, m \geq 0$ *and* $\gamma = 0$ *- for* $m \in [1,2], r \leq \frac{1}{2}, u \geq \frac{1}{2}$ *the problem is regular. In general, whenever* $m \geq 1$ *- i.e. the signal density is convex - for low enough* r, *the problem is regular whenever it is non-trivial, regardless of other parameters.*

D Omitted Proofs

ΔX, as usual, denotes the set of Borel probability measures on any given set X. $Pr(x)$ denotes the probability of an event x. With m_1 and m_0 as defined in (2), we define the following two functions useful for our subsequent analysis, $M_1, M_0 : T \to \mathbb{R}$:

$$M_1(t) = \int_t^1 m_1(t),$$
$$M_0(t) = \int_0^t m_(t). \qquad (M_1, M_0)$$

D.1 Preliminaries

D.1.1 Expression for principal's value

As mentioned in our microfounded model section 5.1, here we derive (Principal's Value).

Since the agent does not know his own quality type, the posterior distribution of the state for any agent receiving a signal t depends on the "average" signal structure of all quality types. We modify the notation for the mapping between signals and beliefs introduced in (1) – μ – to make its dependence on the prior explicit. Specifically, we now define $\mu : T \times [0,1] \to [0,1]$ as:

$$\mu(t;\pi) := Pr(\omega = 1|t) = \frac{\overline{f}_1(t)\pi}{\overline{f}_1(t)\pi + \overline{f}_0(t)(1-\pi)} \qquad (1)$$

The implication of a chosen mechanism $(\widehat{a}_1, \widehat{a}_0) : T \to [0,1]$ the principal cares about is the probability of passing for each learning type v induced by it. Using $Pr(\mathcal{I}|\mathcal{I}')$ to denote the probability of event $\mathcal{I}$ conditional on event $\mathcal{I}'$, where $\mathcal{I}$ and $\mathcal{I}'$ are Borel subsets of $\Omega \times T \times \mathcal{V} \times \{pass, fail\}$, we have:

$$Pr(pass|v) := \int_{t,\omega} Pr(pass|(t,\omega), v) Pr((t,\omega)|v) d\nu(v)$$

Obviously, $Pr(pass|(t,\omega), v) = Pr(pass|(t,\omega)) = \widehat{a}_\omega(t)$. Moreover, $Pr((t,\omega)|v) = Pr(\omega) f(t|\omega, v) dt$ ($\because \omega \perp\!\!\!\perp v$). Combining these, the principal's value from using a mechanism $(\widehat{a}_1, \widehat{a}_0) : [0,1] \to [0,1]$ is given by:

$$\begin{aligned} V(\widehat{a}_1, \widehat{a}_0) &= \int_v v \left(\sum_\omega Pr(\omega) \int_0^1 \widehat{a}_\omega(t) f(t|\omega, v) dt \right) d\nu(v) \\ &= \int_v v \left(\pi \int_0^1 \widehat{a}_1(t) f(t|1,v) dt + (1-\pi) \int_0^1 \widehat{a}_0(t) f(t|0,v) dt \right) d\nu(v) \\ &= \pi \int_0^1 \widehat{a}_1(t) \underbrace{\left(\int_v v f(t|1,v) d\nu(v) \right)}_{=:m_1(t)} dt + (1-\pi) \int_0^1 \widehat{a}_0(t) \underbrace{\left(\int_v v f(t|0,v) d\nu(v) \right)}_{=:m_0(t)} dt \end{aligned} \qquad (2)$$

We can change the order of integrals, as done in the last line, by Fubini's theorem, because each integrand is integrable. In particular, for each $\omega \in \{0,1\}$:

$$\begin{aligned}
&\int_v \int_t |v\widehat{a}_\omega(t) f(t|\omega, v)| dt d\nu(v) \\
&= \int_v \int_t |v| \widehat{a}_\omega(t) f(t|\omega, v) dt d\nu(v) \\
&\leq \int_v \int_t \overline{v} f(t|\omega, v) dt d\nu(v) \\
&= \overline{v}.
\end{aligned}$$

Relabelling m_1 and m_0 as in (2), we get back (Principal's Value). By our boundedness assumptions on $f(\cdot;\cdot,\cdot)$'s and $f'(\cdot;\cdot,\cdot)$'s, m_ω's are well-defined and inherit their continuous differentiability from $f(\cdot;\cdot,\cdot)$'s.

D.1.2 Connection with the problem formulation in terms of beliefs

Below we also formulate the problem in terms of the belief p – a formulation we use for the majority of the main analysis. Note that $p = \mu(t;\pi)$ where $\mu(\cdot;\pi)$ is strictly increasing and differentiable, by our assumption of differentiability of the signal densities. Hence $\mu'(t;\pi) > 0 \ \forall \ t \in [0,1]$. $\therefore \ dt = \frac{dp}{\mu'(\mu^{-1}(p;\pi))}$. With that, we define $a_\omega(p) := \widehat{a}_\omega(\mu^{-1}(p;\pi))$ and $v_\omega(p) := \frac{m_\omega(\mu^{-1}(p;\pi))}{\mu'(\mu^{-1}(p;\pi))}$, $\omega \in \{0,1\}$. Therefore the (scaled) objective function of the principal becomes the following, as in (1).

$$\widehat{V}(a_1, a_0) = \pi \int_{\mu(0;\pi)}^{\mu(1;\pi)} V_1(p) a_1(p) dp + (1-\pi) \int_{\mu(0;\pi)}^{\mu(1;\pi)} V_0(p) a_0(p) dp$$

D.1.3 Principal's value in terms of her virtual value

Using (Integral Formulas) we can express the principal's ex-ante value in terms of its virtual value as follows:

$$
\begin{aligned}
\mathbb{V}\left(a_1, a_0\right) &= \pi \int_p v_1(p) a_1(p) dp + (1-\pi) \int_p v_0(p) a_0(p) dp \\
&= \pi \int_p v_1(p) \left(U(p_0) + \int_{p_0}^{p} q\left(p'\right) dp' - pq(p) \right) dp \\
&+ (1-\pi) \int_p v_0(p) \left(U(p_0) + \int_{p_0}^{p} q\left(p'\right) dp' + (1-p) q(p) \right) dp \\
&= U(p_0)\mathbb{E}v + \int_{p \leq p_0} q(p) \underbrace{\left(\pi(1-p)v^1(p) - (1-\pi)pv^0(p) + \int_{p' \leq p} v(p') dp' \right)}_{=:\chi_L(p)} dp \\
&+ \int_{p \geq p_0} q(p) \underbrace{\left(\pi(1-p)v^1(p) - (1-\pi)pv^0(p) + \int_{p' \geq p} v(p') dp' \right)}_{=:\chi_R(p)} dp \qquad (3) \\
&= U(p_0)\mathbb{E}v + \int_{p \leq p_0} \chi_0(p) q(p) dp + \int_{p \geq p_0} \chi_1(p) q(p) dp \qquad (4)
\end{aligned}
$$

The third equality comes from the usual technique of changing the order of the integrals using Fubini's theorem. The details of the expression in (3) are as follows.

$$
\begin{aligned}
&\int_{p=\underline{p}}^{p=\overline{p}} v^1(p) \left(\int_{x=p_0}^{x=p} q(x) dx \right) dp \\
&= \int_{p=\underline{p}}^{p=\overline{p}} v^1(p) \left(\int_{x=\underline{p}}^{x=\overline{p}} \left(q(x) \left(1(p_0 \leq x \leq p) 1(p \geq p_0) + 1(p_0 \geq x \geq p) 1(p \leq p_0) \right) \right) dx \right) dp \\
&= \int_{p=\underline{p}}^{p=\overline{p}} v^1(p) \left(\int_{x=\underline{p}}^{x=\overline{p}} \left(q(x) \left(1(p \geq \max\{x, p_0\}) 1(x \geq p_0) + 1(p \leq \min\{x, p_0\}) 1(x \leq p_0) \right) \right) dx \right) dp \\
&= \int_{p=\underline{p}}^{p=\overline{p}} v^1(p) \left(\int_{x=\underline{p}}^{x=\overline{p}} \left(q(x) \left(1(p \geq x) 1(x \geq p_0) + 1(p \leq x) 1(x \leq p_0) \right) \right) dx \right) dp \\
&= \int_{p=\underline{p}}^{p=\overline{p}} \int_{x=\underline{p}}^{x=\overline{p}} v^1(p) q(x) 1(p \geq x) 1(x \geq p_0) dx dp + \int_{p=\underline{p}}^{p=\overline{p}} \int_{x=\underline{p}}^{x=\overline{p}} v^1(p) q(x) 1(p \leq x) 1(x \leq p_0) dx dp
\end{aligned}
$$

Now we use the standard technique of changing the order of integrals, which we can do by Fubini's theorem. This gives us the following simplification:

$$\int_{p=\underline{p}}^{p=\overline{p}} v^1(p) \left(\int_{x=p_0}^{x=p} q(x)dx \right) dp$$
$$= \int_{x=\underline{p}}^{x=\overline{p}} q(x)1(x \geq p_0) \left(\int_{p=\underline{p}}^{p=\overline{p}} v^1(p)1(p \geq x)dp \right) dx + \int_{x=\underline{p}}^{x=\overline{p}} q(x)1(x \leq p_0) \left(\int_{p=\underline{p}}^{p=\overline{p}} v^1(p)1(p \leq x)dp \right) dx$$
$$= \int_{x \geq p_0} q(x) \left(\int_{p \geq x} v^1(p)dp \right) dx + \int_{x \leq p_0} q(x) \left(\int_{p \leq x} v^1(p)dp \right) dx$$

D.2 Derivation of the optimal mechanism

In this section we keep the general "base type" (p_0) based formulation introduced in Section C, particularly envelope formula 1. Obviously, all of the analysis goes through if we put $p_0 = \underline{p}$ instead, as done in the main text, Lemma 1.

D.2.1 Proof of Lemma 1

With slight abuse of notation, let $U(p) = a_0(p) + pq(p)$ for all p and $U(p, p') = \widehat{a}_0(p') + pq(p')$, $\forall\ p, p'$.

Necessity. (IC) requires that for each pair of types p, p' both p and p' prefer truth telling:

$$U(p) \geq U(p, p') = U(p') + (p - p')q(p')$$
$$U(p') \geq U(p', p) = U(p) + (p' - p)q(p)$$

Combining:

$$(p - p')q(p) \geq U(p) - U(p') \geq (p - p')q(p')$$

This implies monotonicity.

Assuming $p > p'$ and $p < p'$ we also have, respectively:

$$q(p) \geq \frac{U(p) - U(p')}{p - p'} \geq q(p') \text{ and } q(p) \leq \frac{U(p) - U(p')}{p - p'} \leq q(p')$$

Taking the limit $p' \uparrow p$ we have:

$$U'(p) = q(p)\ \forall\ p \text{ s.t. } U'(p) \text{ exists}$$

From standard arguments we also know $U(p)$ is Lipschitz, therefore absolutely continuous. Therefore for any reference type p_0 we can write the following, giving us the necessity of (1):

$$U(p) = U(p_0) + \int_{p_0}^{p} q(p)dp$$

Sufficiency. Fix any type p. $U(p, p') = U(p') + (p - p')q(p')$. By (1), $U(p) = U(p') + \int_{p'}^{p} q(\widetilde{p})d\widetilde{p}$. Therefore $U(p) - U(p, p') = \int_{p'}^{p}(q(\widetilde{p}) - q(p'))d\widetilde{p}$. Considering cases where $p' > p$ and $p' < p$, and using monotonicity of q, it is easy to see that this implies $U(p) \geq U(p, p')$ for all p'.

D.2.2 Constructing the optimal mechanism

Let us denote $q(\underline{p})$ and $q(\overline{p})$ by $\underline{q}$ and $\overline{q}$ respectively.

Claim D.2.1. *The optimal mechanism is either "pass regardless of type and state" or "do not pass regardless of type and state" or given by the solution to one of the following two programs, whichever has higher value:*

$$\begin{aligned} \max_{q(.),\bar{q},p_0} \mathbb{E}v + \int_{p\leq p_0} (\chi_0(p) + \mathbb{E}v)q(p)dp + \int_{p\geq p_0} \chi_1(p)q(p)dp \\ s.t.\ q \in [-1, \bar{q}] \\ q \text{ non-decreasing} \\ \int_P q(p)dp = \overline{pq} - \underline{pq} - 1 \\ \overline{q} \in [-1, 1]. \end{aligned} \tag{$\overline{P}$}$$

$$\begin{aligned} \max_{q(.),\bar{q},p_0} \mathbb{E}v + \int_{p\leq p_0} \chi_0(p)q(p)dp + \int_{p\geq p_0} (\chi_1(p) - \mathbb{E}v)q(p)dp \\ s.t. \quad q(.) \in [\underline{q}, 1] \\ q \text{ non-decreasing} \\ \int_P q(p)dp = 1 - (1 - \overline{p})\overline{q} + (1 - \underline{p})\underline{q} \\ \underline{q} \in [-1, 1]. \end{aligned} \tag{$\underline{P}$}$$

Proof. The constraints in (Problem) require $\int_{\underline{p}}^{p_0} q\,(p)\,dp - (1 - \underline{p})\underline{q} \leq U(p_0) \leq 1 - \int_{p_0}^{\overline{p}} q\,(p)\,dp - (1 - \overline{p})\overline{q}$ and $\overline{pq} - \int_{p_0}^{\overline{p}} q\,(p)\,dp \leq U(p_0) \leq 1 + \int_{\underline{p}}^{p_0} q(p)dp + \underline{pq}$. Therefore we

need $\int_P q(p)dp \geq \overline{pq} - \underline{pq} - 1$ and $\int_P q(p)dp \leq 1 - (1-\overline{p})\overline{q} + (1-\underline{p})\underline{q}$. Using these, (Problem) reduces to:

$$\max_{p_0, U(p_0), q(.)} U(p_0)\mathbb{E}v + \int_P \chi(p)q(p)dp \tag{1}$$

$$\text{s.t. } q \in [-1,1], q \text{ non - decreasing}$$

$$\overline{pq} - \underline{pq} - 1 \leq \int_P q(p)dp \leq 1 - (1-\overline{p})\overline{q} + (1-\underline{p})\underline{q} \tag{2}$$

$$\max\left\{\int_{\underline{p}}^{p_0} q(p)\,dp - (1-\underline{p})\underline{q}, \overline{pq} - \int_{p_0}^{\overline{p}} q(p)\,dp\right\}$$

$$\leq U(p_0) \leq \min\left\{1 - \int_{p_0}^{\overline{p}} q(p)\,dp - (1-\overline{p})\overline{q}, 1 + \int_{\underline{p}}^{p_0} q(p)dp + \underline{pq}\right\}. \tag{3}$$

□

Proof. Suppose, by way of contradiction, the optimal mechanism is neither constant nor given by the solution to ($\overline{P}$) or ($\underline{P}$). Our proof strategy is to derive a contradiction by considering several cases. We could have set up a Lagrangian and made an argument based on the complementary slackness conditions of the constraints. However we present a direct proof as we believe this is more intuitive.

Before proceeding further we prove a claim below which would be helpful in our analysis of the cases.

Claim D.2.2. *Any solution to* (P1) *is a solution to* (P2)*, where* (P1) *and* (P2) *are as given below.*

$$\begin{aligned} &\max_{p_0, q(.)} \int_P \chi(p)q(p)dp \\ &s.t.\ q \in [-1,1] \\ &\quad q \text{ non-decreasing}. \end{aligned} \tag{P1}$$

$$\begin{aligned} &\max_{\overline{q}, q(.)} \int_P \chi(p)q(p)dp \\ &s.t.\ q \in [-1,\overline{q}] \\ &\quad q \text{ non-decreasing}, \\ &\quad \overline{q} \in [-1,1]. \end{aligned} \tag{P2}$$

Proof. By Proposition 2.1 of Winkler (1988), for any fixed $\overline{q}$, a solution to the problem of maximizing $\int_P \chi(p)q(p)dp$ subject to $q \in [-1,\overline{q}]$ and q non-decreasing, is either $q(p) = -1$ or $q(p) = \overline{q}$ or $q(p) = -1 + (\overline{q}+1)\mathbf{1}(p \geq p^*)$ for some $p^* \in (0,1)$. Therefore the solution to (P2) is given by:

$$\max_{\overline{q} \in [-1,1], p^* \in [0,1]} \left\{0, \overline{q}\int_P \chi(p)dp, \overline{q}\int_{p^*}^{1} \chi(p)dp - \int_{-1}^{p^*} \chi(p)dp\right\}$$

The maximum value of $\overline{q}\int_P \chi(p)dp$ is clearly achieved for $\overline{q} \in \{-1, 1\}$, depending on if $\int_P \chi(p)dp \geq 0$ or $\int_P \chi(p)dp < 0$. If a $p^* \in (0,1)$ is to maximize $\left(\overline{q}\int_{p^*}^{1} \chi(p)dp - \int_{-1}^{p^*} \chi(p)dp\right)$, the first order condition requires that it must satisfy $(\overline{q}+1)\chi(p^*) = 0$, i.e. $\chi(p^*) = 0$ for any $\overline{q} > -1$.

Letting $V_{P2}(\overline{q}) = \max_{p^* \in [0,1]} \left(\overline{q}\int_{p^*}^{1} \chi(p)dp - \int_{-1}^{p^*} \chi(p)dp\right)$, by the envelope theorem, $\frac{\partial V_{P2}(\overline{q})}{\partial \overline{q}} = \int_{p^*}^{1} \chi(p)dp$. Therefore for any $p^* \in \{\chi(p) = 0\}$ such that $\int_{p^*}^{1} \chi(p)dp \geq 0$ (respectively < 0) the optimal $\overline{q} = 1$ (respectively -1).

If the optimal $\overline{q}$ in any of these cases is -1, the optimal q is the constant function $q(p) = -1$, which is feasible for (P1). If the optimal $\overline{q}$ is 1, (P2) boils down to (P1). □

- **Case I: $\mathbb{E}v = 0$.** In this case, in program (1), (3) can be ignored as long as (2) is satisfied. Therefore the problem boils down to:

$$\max_{p_0, q(.)} \int_P \chi(p)q(p)dp \tag{4}$$
$$\text{s.t. } q \in [-1, 1], q \text{ non - decreasing}$$
$$\overline{q} - 1 \leq \int_P q(p)dp \leq \underline{q} + 1. \tag{5}$$

  Suppose none of the constraints (5) binds. Then the solution solves the following "standard" screening problem:

$$\max_{p_0, q(.)} \int_P \chi(p)q(p)dp$$
$$\text{s.t. } q \in [-1, 1], q \text{ non - decreasing}$$

  We know its solution is given by either a constant $q(.)$, or one with a single step, of the form $q(p) = -1 + 2 \times \mathbf{1}(p \geq p^*)$ for some $p^* \in (0,1)$. If it's the latter, $\underline{q} = -1, \overline{q} = 1$. Therefore this solution satisfies (5) – and therefore solves (4) – if and only if $p^* = \frac{1}{2}$. Therefore in this case both the constraints in (5) hold with equality. By Claim D.2.2, in this case $q(p) = -1 + 2 \times \mathbf{1}\left(p \geq \frac{1}{2}\right)$ solves (P2), which is a relaxed version of ($\overline{P}$). But $q(p) = -1 + 2 \times \mathbf{1}\left(p \geq \frac{1}{2}\right)$ satisfies the constraints of ($\overline{P}$), and is therefore a solution to ($\overline{P}$).

  Clearly, if either of the constraints in (5) binds, (4) boils down to either ($\overline{P}$) or ($\underline{P}$).

- **Case II:** $\mathbb{E}v > 0$. In this case, clearly the constraint $U(p_0) \leq \min\left\{1 - \int\limits_{p \geq p_0} q(p)dp, 1 + \int\limits_{p \leq p_0} q(p)dp\right\}$ in (3) must bind. Therefore the problem boils down to:

$$\max_{p_0, q(.)} \min\left\{1 - \int\limits_{p \geq p_0} q(p)dp, 1 + \int\limits_{p \leq p_0} q(p)dp\right\} + \int_P \chi(p)q(p)dp$$
$$\text{s.t. } q \in [-1, 1], q \text{ non - decreasing}$$
$$\overline{q} - 1 \leq \int\limits_P q(p)dp \leq \underline{q} + 1.$$

$1 + \int\limits_{p \leq p_0} q(p)dp \gtreqless 1 - \int\limits_{p \geq p_0} q(p)dp \iff \int\limits_P q(p)dp \gtreqless 0$. Hence the solution to the above program is given by the solution to one of the following programs – one assuming $\int\limits_P q(p)dp \geq 0$ and the other $\int\limits_P q(p)dp \geq 0$ – whichever gives higher value.

$$\mathbb{E}v + \max_{p_0, q(.)} \int_P (\chi(p) - \mathbb{E}v(p \geq p_0))q(p)dp \qquad \mathbb{E}v + \max_{p_0, q(.)} \int_P (\chi(p) + \mathbb{E}v(p \leq p_0))q(p)dp$$
$$\text{s.t. } q \in [-1, 1] \qquad (P_1^+) \qquad \text{s.t. } q \in [-1, 1] \qquad (P_2^+)$$
$$q \text{ non-decreasing}, \qquad q \text{ non-decreasing},$$
$$\overline{q} - 1 \leq \int\limits_P q(p)dp \leq 0. \qquad (6) \qquad 0 \leq \int\limits_P q(p)dp \leq \underline{q} + 1. \qquad (7)$$

Suppose the solution is given by the solution to (P_1^+).

Suppose, fist, that no side of the constraint (6) binds. The only way such a solution can satisfy (6) is if it is either constant or given by $q(p) = -1 + 2 \times \mathbf{1}\left(p \geq \frac{1}{2}\right)$. In the latter case, by arguments similar to those in Case I, $q(p) = -1 + 2 \times \mathbf{1}\left(p \geq \frac{1}{2}\right)$ is a solution to $(\overline{P})$.

Note that both sides of (6) cannot bind (though both can hold with equality, of course.)

Next, suppose the constraint $\overline{q} - 1 \leq \int\limits_P q(p)dp$ binds. Then (P_1^+) boils down to $(\overline{P})$.

Finally, suppose the constraint $\int\limits_P q(p)dp \leq 0$ binds. In that case (P_1^+) is equivalent to the following relaxed program:

$$\begin{aligned}
&\mathbb{E}v + \max_{p_0, q(.)} \int_P (\chi(p) - \mathbb{E}v(p \geq p_0))q(p)dp \\
&\text{s.t. } q \in [-1, 1] \qquad (P^+_{1,R}) \\
&q \text{ non-decreasing }, \\
&\int_P q(p)dp \leq 0. \qquad (8)
\end{aligned}$$

Consider the following further relaxed problem:

$$\begin{aligned}
&\mathbb{E}v + \max_{p_0, q(.)} \int_P (\chi(p) - \mathbb{E}v(p \geq p_0))q(p)dp \\
&\text{s.t. } q \in [-1, 1] \qquad (P^+_{1,RR}) \\
&q \text{ non-decreasing }.
\end{aligned}$$

$\int_P q(p)dp \leq 0$ in (P^+_1) binds only if the solution to $(P^+_{1,RR})$ is $q_{1,R}(p) = -1 + 2 \times \mathbf{1}\left(p \geq p^*_{1,R}\right)$ for some $p^*_{1,R} \in \left[\underline{p}, \frac{\underline{p}+\overline{p}}{2}\right)$. Therefore any solution to $(P^+_{1,R})$ – say $q^*_{1,R}$ – is a convex combination of $q_{1,R}$ and an extreme point of the feasible set of $(P^+_{1,RR})$ – say $q_{1,RR}$ – which remains feasible with the constraint (8) (Needs linear programming citation). Hence we must have $q_{1,RR}(p) = -1 + 2 \times \mathbf{1}\left(p \geq p^*_{1,RR}\right)$ for some $p^*_{1,RR} \in \left[\frac{\underline{p}+\overline{p}}{2}, \overline{p}\right]$.

If $p^*_{1,R} = \underline{p}$ and $p^*_{1,RR} = \overline{p}$, the solution to $(P^+_{1,R})$ and – therefore (P^+_1) – must be the equally weighted convex combination of $q_{1,R}$ and $q_{1,RR}$, i.e. the constant function $q^*_{1,R}(p) = 0$.

If $p^*_{1,R} \in \left(\underline{p}, \frac{\underline{p}+\overline{p}}{2}\right)$ **or** $p^*_{1,RR} \in \left[\frac{\underline{p}+\overline{p}}{2}, \overline{p}\right)$, the convex combination $q^*_{1,R}$ must have either $q^*_{1,R}(1) = 1$ or $q^*_{1,R}(0) = -1$.

Note that $(P^+_{1,R})$ is a relaxed version of $(\overline{P})$, because $\int_P q(p)dp = \bar{q} - 1$ and $\bar{q} \leq 1 \implies \int_P q(p)dp \leq 0$. Therefore $q^*_{1,R}$ is not a solution to $(\overline{P})$ only if it does not satisfy the constraint $\int_P q(p)dp = \bar{q} - 1$, i.e. only if $\bar{q} < 1$. (Recall that $\int_P q^*_{1,R}(p)dp = 0$ by assumption.) In this case we must have $q^*_{1,R}(0) = -1$.
Under our assumptions, (P^+_1) is equivalent to the following program:

$$\begin{aligned}
&\mathbb{E}v + \max_{p_0, q(.)} \int_P (\chi(p) - \mathbb{E}v(p \geq p_0))q(p)dp \\
&\text{s.t. } q \in [-1, 1] \qquad (P^+_0) \\
&q \text{ non-decreasing }, \\
&\int_P q(p)dp = 0. \qquad (9)
\end{aligned}$$

Recall that the value of (P_0^+) is weakly higher than that of (P_2^+), by assumption. But (P_0^+) is (P_2^+) with the constraint $\int_P q(p)dp \geq 0$ binding. Therefore any solution to (P_0^+) must be a solution to (P_2^+) as well. Any solution to $(\underline{P})$ is clearly a solution to (P_2^+) with the additional constraint $\int_P q(p)dp \leq \underline{q} + 1$ - i.e. with one side of (7) binding. But the value of (P_2^+) is given by the value of (P_0^+), as we just argued. Therefore at any solution to $(\underline{P})$ we must have $\int_P q(p)dp = 0$, i.e. $\underline{q} = -1$. Therefore if $q^*_{1,R} < 1$, the solution to (P_1^+) and therefore (**??**) is given by the solution to $(\underline{P})$.

Almost identical arguments show that if the solution is given by the solution to (P_2^+), it is also either constant or given by the solution to either $(\overline{P})$ or $(\underline{P})$.

- **Case III:** $\mathbb{E}v < 0$. Analogously as in Case II, in this case the constraint $\max\left\{-\underline{q} + \int_{p\leq p_0} q(p)dp, \overline{q} - \int_{p\geq p_0} q(p)dp\right\} \leq U(p_0)$ in (3) must bind.
  Therefore the problem boils down to:

$$\max_{p_0, q(.)} \max\left\{-\underline{q} + \int_{p\leq p_0} q(p)dp, \overline{q} - \int_{p\geq p_0} q(p)dp\right\} \mathbb{E}v + \int_P \chi(p)q(p)dp \tag{10}$$

$$\text{s.t. } q \in [-1,1], q \text{ non - decreasing}$$

$$\overline{q} - 1 \leq \int_P q(p)dp \leq \underline{q} + 1. \tag{11}$$

Note that in this case the objective is convex in q. Therefore the maximum cannot be attained in the interior of the feasible region, and hence constraints which do not bind at the optimum can be removed.

Suppose the constraint $\int_P q(p)dp \leq \underline{q} + 1$ binds at the optimum (similarly as in Case II, both sides of (11) cannot bind). In this case the problem boils down to:

$$\max_{q(.),\underline{q}} -\underline{q}\mathbb{E}v + \int_P (\chi(p) + \mathbb{E}v(p \leq p_0))q(p)dp$$

$$\text{s.t.} \quad q(.) \in [\underline{q}, 1]$$

$$q \text{ non-decreasing}$$

$$\int_P q(p)dp = \underline{q} + 1$$

$$\underline{q} \in [-1, 1].$$

Replacing $-\underline{q}$ in the objective with $1-\int_P q(p)dp$ – using the equality constraint – gives us ($\underline{P}$). Analogous reasoning shows that when, in the solution to (10), the constraint $\overline{q}-1 \leq \int\limits_P q(p)dp$ binds instead, the problem boils down to ($\overline{P}$).

Suppose none of the constraints bind. Using the same extreme points based reasoning as before, this can happen only if the solution is a constant q or $q(p) = -1 + 2 \times \mathbf{1}\left(p \geq \frac{1}{2}\right)$. In the latter case the solution is a solution to the following problem:

$$\begin{aligned}
&\max_{p_0, q(.)} \mathbb{E}v \left(1 - \int\limits_{p \geq p_0} q(p)dp\right) + \int_P \chi(p)q(p)dp \\
&\text{s.t. } q \in [-1, 1] \\
&q \text{ non-decreasing }, \\
&\int\limits_P q(p)dp = 0.
\end{aligned}$$

Similarly as in Case II, in this case both the constraints in (11) hold with equality at the optimum, so the solution is given by the solution to both of ($\overline{P}$) and ($\underline{P}$).

□

Proof of Proposition B.1. By Claim D.2.1 we know that when the optimal mechanism is not constant, it is given by the solution to either ($\overline{P}$) or ($\underline{P}$). The proof strategy is as follows: We first show that the solution could feature at most two thresholds and unless the solutions to ($\overline{P}$) and ($\underline{P}$) coincide, the solution to either is given by a single threshold q. If the solution solves ($\overline{P}$), this threshold must be below one half, and if it solves ($\underline{P}$) it must be above one half. These correspond to the first and third cases in (1) respectively. The solution could feature two thresholds only if the solutions to ($\overline{P}$) and ($\underline{P}$) coincide, $\int\limits_P q(p)dp = 0$ and these thresholds average to one half. This corresponds to the second case in (1).

Claim D.2.3. *The solution q to* (Problem) *is a step function featuring at most two steps.*

Proof. By Claim D.2.1, the solution is either constant – in which case it is a step function featuring zero steps – or given by the solution to ($\overline{P}$) or ($\underline{P}$).

Consider ($\overline{P}$). For any fixed $\overline{q} \in (-1, 1]$, ($\overline{P}$) is a linear programs in q, with the constraint that q is non-decreasing, and one additional equality constraint. By Proposition 2.1, Winkler (1988), all extreme points of the set $S = \{q \in [0, \overline{q}]^P : q \text{ non-decreasing}\}$ are of the form $q(p) = -1 + (\overline{q}+1)\mathbf{1}(p \geq p^*)$ for some $p^* \in [0, 1]$.

Therefore those of the set $F = \{q \in [0,\overline{q}]^P : q \text{ non-decreasing}, \int_P q(p)dp = \overline{q} - 1\}$ - which is the subset of S satisfying that one additional linear constraint – must be given by the (possibly degenerate) convex combination of two extreme points of S. In case the solution q^* to $(\overline{P})$ is given by an extreme point of F which is a proper convex combination of two extreme points of S, say $q_i(p) = -1 + (\overline{q}+1)\mathbf{1}(p \geq p_i^*), i \in \{1,2\}$ such that $p_i^* \in (0,1)$ for $i \in \{1,2\}$, q^* features two steps. □

D.3 Proof of Proposition 1: The two characterizations of regularity

Claim D.3.1. *The problem is regular if and only if, for all signal realizations, $t \in [0,1]$, the following holds:*

$$m_1(t)\left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)}\right) > m_0(t)\phi_1(t)\left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_0''(t)}{\phi_0'(t)}\right) \tag{Reg}$$

Proof. The problem is regular – i.e. $\chi(p)$ is increasing iff $\chi'(p) > 0$. Hence we need,

$$\begin{aligned}
&\pi((1-p)v^{1\prime}(p) - 2v^1(p)) - (1-\pi)(pv^{0\prime}(p) + 2v^0(p)) > 0 \\
&\implies \pi v^1(p)\left((1-p)\frac{v^{1\prime}(p)}{v^1(p)} - 2\right) - (1-\pi)v^0(p)\left(p\frac{v^{0\prime}(p)}{v^0(p)} + 2\right) > 0 \qquad (1)
\end{aligned}$$

As defined earlier, $v^0(p) = \frac{m_0(\mu^{-1}(p;\pi))}{\mu'(\mu^{-1}(p;\pi))}$. Let t be such that $\mu(t) = p$. Recall that t is unique for any given p, owing to strict increasingness of μ. Differentiating v^0 and some algebra shows that:

$$\begin{aligned}
&\frac{v^{0\prime}(p)}{v^0(p)} = \left(\frac{m_0'(\mu^{-1}(p;\pi))}{m_0(\mu^{-1}(p;\pi))} - \frac{\mu''(\mu^{-1}(p;\pi))}{\mu'(\mu^{-1}(p;\pi))}\right)\left(\frac{1}{\mu'(\mu^{-1}(p;\pi))}\right) \\
&\iff \frac{pv^{0\prime}(p)}{v^0(p)} = \left(\frac{m_0'(t)}{m_0(t)} - \frac{\mu''(t)}{\mu'(t)}\right)\left(\frac{\mu(t)}{\mu'(t)}\right) \qquad (2)
\end{aligned}$$

Let $\mu_c(t) := 1 - \mu(t)$ for all t. Therefore $1 - p = \mu_c(t)$. Then, analogously as above we have:

$$\frac{(1-p)v^{1\prime}(p)}{v^1(p)} = \left(\frac{m_1'(t)}{m_1(t)} - \frac{\mu_c''(t)}{\mu_c'(t)}\right)\left(-\frac{\mu_c(t)}{\mu_c'(t)}\right) \tag{3}$$

Hence, recalling that $v^\omega(p) = \frac{m_\omega(t)}{\mu'(t)}$, where $p = \mu(t), \omega \in \{0,1\}$, by (1) we need,

$$\begin{aligned}
&\pi\frac{m_1(t)}{\mu'(t)}\left[\left(\frac{m_1'(t)}{m_1(t)} - \frac{\mu_c''(t)}{\mu_c'(t)}\right)\left(-\frac{\mu_c(t)}{\mu_c'(t)}\right) - 2\right] > (1-\pi)\frac{m_0(t)}{\mu'(t)}\left[\left(\frac{m_0'(t)}{m_0(t)} - \frac{\mu''(t)}{\mu'(t)}\right)\left(\frac{\mu(t)}{\mu'(t)}\right) + 2\right] \\
&\implies \pi\left(-\frac{\mu_c(t)m_1(t)}{\mu_c'(t)}\right)\left[\frac{m_1'(t)}{m_1(t)} - \frac{\mu_c''(t)}{\mu_c'(t)} + 2\frac{\mu_c'(t)}{\mu_c(t)}\right] > (1-\pi)\left(\frac{\mu(t)m_0(t)}{\mu'(t)}\right)\left[\frac{m_0'(t)}{m_0(t)} - \frac{\mu''(t)}{\mu'(t)} + 2\frac{\mu'(t)}{\mu(t)}\right.
\end{aligned}$$

Let $\phi_0(t) = \frac{\bar{f}_0(t)}{\bar{f}_1(t)}$. $\therefore \quad \mu(t) = \frac{1}{1+\left(\frac{1-\pi}{\pi}\right)\phi_0(t)}$. Differentiating both sides of that last equation twice we get, $\frac{\mu''(t)}{\mu'(t)} - 2\frac{\mu'(t)}{\mu(t)} = \frac{\phi_0''(t)}{\phi_0'(t)}$. Similarly, letting $\phi_1(t) = \frac{1}{\phi_0(t)}$, $\frac{\mu_c''(t)}{\mu_c'(t)} - 2\frac{\mu_c'(t)}{\mu_c(t)} = \frac{\phi_1''(t)}{\phi_1'(t)}$. Hence - also recalling that $-\mu_c'(t) = \mu'(t)$- the above is equivalent to:

$$m_1(t)\left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)}\right) > m_0(t)\left(\frac{1-\pi}{\pi}\right)\left(\frac{\mu(t)}{\mu_c(t)}\right)\left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_0''(t)}{\phi_0'(t)}\right)$$

Note that $\left(\frac{1-\pi}{\pi}\right)\left(\frac{\mu(t)}{\mu_c(t)}\right) = \phi_1(t)$. Using this in the above expression we have the desired result. □

Claim D.3.2. *V_1 is concave if and only if (Reg) holds.*

Proof. Clearly,

$$\frac{\mathrm{d}V_1}{\mathrm{d}s_1} = \frac{\frac{\mathrm{d}V_1}{\mathrm{d}t}}{\frac{\mathrm{d}s_1}{\mathrm{d}t}}$$

Differentiating both sides of the above equation w.r.t. s_1 we have:

$$\frac{\mathrm{d}^2V_1}{\mathrm{d}s_1^2} = \frac{\frac{\mathrm{d}^2V_1}{\mathrm{d}s_1\mathrm{d}t}}{\frac{\mathrm{d}s_1}{\mathrm{d}t}}$$

From the above equation the numerator of $\frac{\mathrm{d}^2V_1}{\mathrm{d}s_1^2}$ is $\left(\phi_1'(t)\frac{\mathrm{d}^2V_1(\phi_1(t))}{\mathrm{d}^2t} - \phi_1''(t)\frac{\mathrm{d}V_1(\phi_1(t))}{\mathrm{d}t}\right)$. It's denominator is strictly positive. Therefore V_1 is concave in s_1 if and only if:

$$\left(\phi_1'(t)\frac{\mathrm{d}^2V_1(\phi_1(t))}{\mathrm{d}^2t} - \phi_1''(t)\frac{\mathrm{d}V_1(\phi_1(t))}{\mathrm{d}t}\right) > 0 \;\forall\, t \tag{4}$$

From the main text,

$$V_1(\phi_1(t)) = \pi\left(\phi_1(t)\left(\int_0^t m_0(t)dt - v_0^+\right) + \int_t^1 m_1(t)dt\right) + (1-\pi)v_0^+$$

Differentiating the above twice w.r.t. t and using the expressions of $\frac{\mathrm{d}V_1(\phi_1(t))}{\mathrm{d}t}$ and $\frac{\mathrm{d}^2V_1(\phi_1(t))}{\mathrm{d}^2t}$ in (4) we get back (Reg). □

Claim D.3.3. *V_1 is concave if and only if V_0 is concave.*

Proof. Using equations (1) and (2) and analogous expressions for the principal's value from single-threshold tests as a function of threshold when the undistorted state is 0, we have, after some algebra:

$$V_0(s_0) - \pi V_0^+ = \left(\frac{1-\pi}{\pi}\right) s_0 \left(V_1\left(\frac{1}{s_0}\right) - (1-\pi)v_0^+\right) \tag{5}$$

Differentiating both sides twice w.r.t. s_0 we have:

$$v_0''(s_0) = \left(\frac{1-\pi}{\pi}\right)\left(\frac{1}{s_0^3}\right) v_1''\left(\frac{1}{s_0}\right) \tag{6}$$

Clearly, $v_0''(s_0) \gtreqless 0 \iff v_1''\left(\frac{1}{s_0}\right) \gtreqless 0$, which establishes the claim.

□

D.4 Proofs for Section 5.2.1, Main characterization

Note that under any pass-if-correct or fail-if-incorrect test, as defined in the main text, the passing rate in at least one of the states is undistorted on the intensive margin, i.e. takes values only in $\{0, 1\}$. Depending on which state's passing rate is undistorted in this sense, we further divide pass-if-correct or fail-if-incorrect tests into two categories each, and give the following short names to them, for ease of notation.

- F_0: A fail-if-incorrect test with state 0 undistorted (F stands for *fail*, 0 in the subscript captures the undistorted state, and so on for the following tests.)
- F_1: A fail-if-incorrect test with state 1 undistorted
- P_0: A pass-if-correct test with state 0 undistorted
- P_1: A pass-if-correct test with state 1 undistorted

Going forward, we refer the above four categories as *types* of single threshold tests. The following result would help us further consolidate these categories into just two.

Claim D.4.1. *Pass-if-correct (respectively fail-if-incorrect) tests are optimal only if the market is lemon-dropping (respectively cherry-picking).*

Proof. The value from the F and P tests of a given signal threshold are as follows:

$$\begin{aligned} V_-^F(t) &= \pi \int_t^1 m_1(t')dt' + (1-\pi)\left(\frac{\mu(t;\pi)}{1-\mu(t;\pi)}\right)\int_0^t m_0(t')dt' \\ &= \pi\left(\int_t^1 m_1(t')dt' + \phi_1(t)\int_0^t m_0(t')dt'\right) \end{aligned} \tag{1}$$

$$V_-^P(t) = \pi \int_t^1 m_1(t')dt' + (1-\pi)\left[\int_0^t m_0(t')dt' + \left(1 - \frac{\mu(t;\pi)}{1-\mu(t;\pi)}\right)\int_t^1 m_0(t')dt'\right]$$
$$= \pi \int_t^1 \left(m_1(t') - \phi_1(t)m_0(t')\right)dt' + (1-\pi)\int_0^1 m_0(t')dt' \tag{2}$$

Comparing the two with the same threshold,

$$V_-^P(t) \geq V_-^F(t)$$
$$\iff \pi\left(\int_t^1 m_1(t')dt' + \phi_1(t)\int_0^t m_0(t')dt'\right) \geq \pi\int_t^1 \left(m_1(t') - \phi_1(t)m_0(t')\right)dt' + (1-\pi)\int_0^1 m_0(t')dt'$$
$$\iff \pi\phi(t)\int_0^1 m_0(t')dt' \geq (1-\pi)\int_0^1 m_0(t')dt'$$

Now, $\int_0^1 m_0(t')dt' = V_0$. Therefore if we are in a lemon-dropping market, i.e. $\alpha < 1$, $V_-^P(t) \geq V_-^F(t) \iff \phi_1(t) \geq \frac{1-\pi}{\pi}$, which is impossible over the range of t for which the belief is below one half, unless this inequality holds with equality, i.e. the belief-threshold is one half and the pass-if-correct and fail-if-incorrect tests coincide.

On the other hand if we are in a cherry-picking market, i.e. $V_0 < 0$ $V_-^P(t) \geq V_-^F(t) \iff \phi_1(t) \leq \frac{1-\pi}{\pi}$, which holds for every t in the domain of V_-^P and V_-^F.

The proof is similar for F_+ and P_+ tests. □

Claim D.4.2. *Suppose the problem is regular. The principal's maximized value is given by the following:*

In a cherry-picking market: $\mathbb{V} = \max\left\{\max_{\left\{t:\phi_1(t)\leq\frac{1-\pi}{\pi}\right\}} V_-^F(t), \max_{\left\{t:\phi_1(t)\geq\frac{1-\pi}{\pi}\right\}} V_+^F(t), 0\right\}$

In a lemon-dropping market: $\mathbb{V} = \max\left\{\max_{\left\{t:\phi_1(t)\leq\frac{1-\pi}{\pi}\right\}} V_-^P(t), \max_{\left\{t:\phi_1(t)\geq\frac{1-\pi}{\pi}\right\}} V_+^P(t), V_0\right\}$

Proof. If the problem is regular, optimal tests have a single threshold. The expressions then follow from the fact that $V_-^F(t)$ gives the principal's value from a fail-if-incorrect test with threshold t only if t is such that the agent's belief at t is below one half, and similarly for the other expressions. Note that setting $\phi_1(t) = \frac{1-\pi}{\pi}$ gives the simple T-F test, which is always feasible. Finally, if the maximized value from all single-threshold tests falls below the principal's expected value without screening - which is 0 in a cherry-picking market and V_0 in a lemon-dropping market, she chooses not to screen. Putting these together we get the above expressions. □

Claim D.4.3. *Under regularity, V_-^F, V_+^F, V_-^P and V_+^P are qausiconcave.*

Proof. We show first show V_-^F is qausiconcave by showing that ${V^F}'_-(t) < 0$ whenever ${V^F}'_-(t) = 0$.

$$
\begin{aligned}
{V^F}'_-(t) = 0 &\implies -m_1(t) + \phi_1(t)m_0(t) + \phi_1'(t) \int_0^t m_0(t')dt' = 0 \\
&\implies \int_0^t m_0(t')dt' = \frac{m_1(t) - \phi_1(t)m_0(t)}{\phi_1'(t)} \qquad (3)
\end{aligned}
$$

$$
\begin{aligned}
{V^F}''_-(t) &= -m_1'(t) + \phi_1(t)m_0'(t) + 2\phi_1'(t)m_0(t) + \phi_1''(t) \int_0^t m_0(t')dt' \\
&= -m_1'(t) + \phi_1(t)m_0'(t) + 2\phi_1'(t)m_0(t) + \phi_1''(t) \left(\frac{m_1(t) - \phi_1(t)m_0(t)}{\phi_1'(t)} \right) \\
&= \phi_1(t)m_0(t) \left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_1''(t)}{\phi_1'(t)} + 2\frac{\phi_1'(t)}{\phi_1(t)} \right) - m_1(t) \left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)} \right)
\end{aligned}
$$

Note that $\phi_0(t) = \frac{1}{\phi_1(t)}, \therefore \ \frac{\phi_0''(t)}{\phi_0'(t)} = \frac{\phi_1''(t)}{\phi_1'(t)} - 2\frac{\phi_1'(t)}{\phi_1(t)}$. Using this in the above expression we have,

$$
{V^F}''_-(t) = \phi_1(t)m_0(t) \left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_0''(t)}{\phi_0'(t)} \right) - m_1(t) \left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)} \right) \qquad (4)
$$

The above expression is negative by regularity. Therefore V_-^F cannot have a local minima. Therefore it can have at most a single local maxima, which must then be its global maxima. In other words, V_-^F is qausiconcave.

Now we show that V_-^P is quasiconcave as well.

$$
\begin{aligned}
{V^P}'_-(t) = 0 &\implies -m_1(t) + \phi_1(t)m_0(t) - \phi_1'(t) \int_t^1 m_0(t')dt' = 0 \\
&\implies \int_t^1 m_0(t')dt' = -\frac{m_1(t) - \phi_1(t)m_0(t)}{\phi_1'(t)} \qquad (5)
\end{aligned}
$$

$$
\begin{aligned}
{V^P}''_-(t) &= -m_1'(t) + \phi_1(t)m_0'(t) + 2\phi_1'(t)m_0(t) - \phi_1''(t) \int_t^1 m_0(t')dt' \\
&= -m_1'(t) + \phi_1(t)m_0'(t) + 2\phi_1'(t)m_0(t) + \phi_1''(t) \left(\frac{m_1(t) - \phi_1(t)m_0(t)}{\phi_1'(t)} \right) \\
&= \phi_1(t)m_0(t) \left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_0''(t)}{\phi_0'(t)} \right) - m_1(t) \left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)} \right)
\end{aligned}
$$

The last line, again, comes from the fact that $\frac{\phi_0''(t)}{\phi_0'(t)} = \frac{\phi_1''(t)}{\phi_1'(t)} - 2\frac{\phi_1'(t)}{\phi_1(t)}$. The above expression is the same as $V^{F\prime\prime}{}_{-}(t)$ (The t is, of course, not the same as in that case, as the derivatives vanish at different points), which is negative by regularity.

Now we show that V_+^F is qausiconcave as well.

$$V_+^F(t) = (1-\pi)\left(\phi_0(t)\int_t^1 m_1(t')dt' + \int_0^t m_0(t')dt'\right) \tag{6}$$

$$V^{F\prime}{}_{+}(t) = 0 \implies m_0(t) - \phi_0(t)m_1(t) + \phi_0'(t)\int_t^1 m_1(t')dt' = 0$$

$$\implies \int_t^1 m_1(t')dt' = -\frac{m_0(t) - \phi_0(t)m_1(t)}{\phi_0'(t)} \tag{7}$$

$$\begin{aligned}
V^{F\prime\prime}{}_{+}(t) &= m_0'(t) - \phi_0(t)m_1'(t) - 2\phi_0'(t)m_1(t) + \phi_0''(t)\int_t^1 m_1(t')dt' \\
&= m_0'(t) - \phi_0(t)m_1'(t) - 2\phi_0'(t)m_1(t) + \phi_0''(t)\left(-\frac{m_0(t) - \phi_0(t)m_1(t)}{\phi_0'(t)}\right) \\
&= \phi_1(t)m_0(t)\left(\frac{m_0'(t)}{m_0(t)} - \frac{\phi_0''(t)}{\phi_0'(t)}\right) - m_1(t)\left(\frac{m_1'(t)}{m_1(t)} - \frac{\phi_1''(t)}{\phi_1'(t)}\right)
\end{aligned}$$

The second line comes from substituting (7). The third line, comes from the fact that $\frac{\phi_1''(t)}{\phi_1'(t)} = \frac{\phi_0''(t)}{\phi_0'(t)} - 2\frac{\phi_0'(t)}{\phi_0(t)}$. The above expression is the same as $V^{F\prime\prime}{}_{+}(t)$ (The t is, of course, not the same as in that case, as the derivatives vanish at different points), which is negative by regularity. The proof for V_+^P is similar. □

Claim D.4.4. *Either the optimal mechanism is to always or never pass, or there exist $t_j^i \in (0,1)$ such that $V_j^i(t)$ is maximized at $t_j^i, i \in \{P,F\}, j \in \{+,-\}$.*

Proof. From the expressions for $V^{F\prime}{}_{j}(t)$ above, $j \in \{+,-\}$, it is clear that for low enough t it is positive, and for $t = 1$ it is negative. Hence by continuity of $V^{F\prime}{}_{j}(t)$ – which follows from our assumption of continuous differentiability of the signal densities – there exists $t_j^F \in (0,1)$ such that $V^{F\prime}{}_{j}(t) = 0$.

Now consider $V^{P\prime}{}_{-}(t)$. Note that $V^{P\prime}{}_{-}(1) < 0$. Suppose there does not exist $t_-^P \in (0,1)$ such that $V^{P\prime}{}_{-}(t) = 0$. Hence $V^{P\prime}{}_{-}(t) < 0$ for all $t \in (0,1)$. Hence the optimal test of type P_- has a signal threshold of $t = 0$. Under this test there is no screening – everyone is passd in state 1 and everyone is passd with probability $\left(1 - \left(\frac{\pi}{1-\pi}\right)\phi_1(0)\right)$ in state 0. But because P_- can be optimal only if $V_0 \leq 0$, in this case the principal is weakly better off passing everyone with probability one in both states. Hence in this case the optimal test is to never pass. Similarly it can be shown that the same is true for the test of type P_+. □

Claim D.4.5. *Unless the optimal mechanism is to always or never pass, in a cherry-picking market $t_-^F > t_+^F$ and in a lemon-dropping market $t_-^P > t_+^P$, where t_j^i's are as defined in Claim D.4.4, $i \in \{P, F\}, j \in \{+, -\}$.*

Proof. We prove the lemma first for the case when $V_0 \leq 0$.

Suppose the test is not constant for some π. By Claim D.4.2, the maximized value of the principal when using a non-constant test, $\mathbb{V} \leq \max\{V_-^F(t_-^F), V_+^F(t_+^F)\}$. Hence, if the optimal test is not constant, $\max\{V_-^F(t_-^F), V_+^F(t_+^F)\} > 0$. Note that:

$$V_+^F(t) = \left(\frac{1-\pi}{\pi}\right)\phi_0(t)V_-^F(t)$$

Since $\phi_0(t) > 0$ for all t, $V_-^F(t_-^F) > 0 \implies V_+^F(t_-^F) > 0$, $\therefore$ $V_+^F(t_+^F) \geq V_+^F(t_-^F) > 0$. Similarly, $V_+^F(t_+^F) > 0 \implies V_-^F(t_-^F) > 0$. i.e. If the optimal test is not constant, $\min\{V_-^F(t_-^F), V_+^F(t_+^F)\} > 0$.

${V^F}'_+(t) = \left(\frac{1-\pi}{\pi}\right)[\phi_0(t){V^F}'_-(t) + \phi_0'(t)V_-^F(t)]$. Recall that ${V^F}'_-(t_-^F) = 0$, $\phi_0'(t) < 0$ for all t and $V_-^F(t_-^F) > 0$. Therefore ${V^F}'_+(t_-^F) = \left(\frac{1-\pi}{\pi}\right)\phi_0'(t)V_-^F(t_-^F) < 0$. By single-peakedness of V_+^F, as shown in Claim D.4.3, this means $t_+^F < t_-^F$.

Now we consider the case $V_0 > 0$. Some algebra shows:

$$V_+^P(t) - \pi V_0 = \left(\frac{1-\pi}{\pi}\right)\phi_0(t)\left(V_-^P(t) - (1-\pi)V_0\right) \tag{8}$$

At any t, $V_+^P(t) > \pi V_0 \iff V_-^P(t) > (1-\pi)V_0$. Since the problem is regular, the optimal test is not constant for some π and $\alpha < 1$, this must hold for $t \in \{t_+^P, t_-^P\}$.

Differentiating both sides of (8),

$$\frac{{V^P}'_+(t)}{1-\pi} = \frac{1}{\pi}\left(\phi_0(t){V^P}'_-(t) + (V_-^P(t) - (1-\pi)V_0)\phi_0'(t)\right)$$

At $t = t_-^P$ we know $V_-^P(t) - (1-\pi)V_0 > 0$ and ${V^P}'_-(t) = 0$. $\phi_0' < 0$ and $\phi_0 > 0$, so ${V^P}'_+(t_-^P) < 0$. Hence, by quasiconcavity of V_+^P, $t_+^P < t_-^P$. □

Define the following:

$$\underline{\underline{\pi}} := \frac{1}{\phi_1(1)+1} \text{ and } \overline{\overline{\pi}} := \frac{1}{\phi_1(0)+1} \tag{9}$$

Claim D.4.6. *For $\pi \in (0, \underline{\underline{\pi}}]$ (respectively, $\pi \in [\overline{\overline{\pi}}, 1)$) the optimal test is of type F_- (respectively, F_+), regardless of regularity.*

Proof. For $\pi < \underline{\underline{\pi}}$, F_+ tests are infeasible, since $\phi_1(t) < \frac{1-\pi}{\pi}$ for all t. Therefore for $\pi \in (0, \underline{\underline{\pi}}]$ the optimal test is of type F_-, since, as argued in the proof of Claim D.4.2, if it is not constant, $V_-^F(t_-^F) > 0$. Clearly, this does not depend on regularity, because for $\pi < \underline{\underline{\pi}}$, all beliefs are strictly below one half, and hence no other type of test – including those with two thresholds – are feasible. (By Theorem B.1, a two

threshold optimal test must have exactly one belief threshold on each side of one half.)

Similarly the bracketed parts follow. □

Completing the proof. Again, we provide the proof for $V_0 \leq 0$. The proof for $V_0 > 0$ is identical. Since we consider $V_0 \leq 0$ and assume regularity, by Proposition 2 the only types of tests we need to consider are F_- and F_+.

Suppose the there exists some $\pi \in (0,1)$ such that the optimal test is not constant for π.

Note that $\left\{t : \phi_1(t) \geq \frac{1-\pi}{\pi}\right\}$ is an interval, by strict increasingness of ϕ_1. By single-peakedness of V_+^F as shown by Claim D.4.3, for $\pi \in \left[\underline{\pi}, \frac{1}{\phi_1(t_+^F)+1}\right]$, V_+^F is strictly decreasing in t over $\left\{t : \phi_1(t) \geq \frac{1-\pi}{\pi}\right\}$. By Claim D.4.5, $t_+^F < t_-^F$. Hence the same holds for $\pi \in \left[\underline{\pi}, \frac{1}{\phi_1(t_-^F)+1}\right] \subsetneq \left[\underline{\pi}, \frac{1}{\phi_1(t_+^F)+1}\right]$. Therefore $\max_{\left\{t:\phi_1(t)\geq\frac{1-\pi}{\pi}\right\}} V_+^F(t) = V_+^F(\widehat{t}(\pi)) = \pi \int_{\widehat{t}(\pi)}^{1} m_1(t')dt' + (1-\pi)\int_0^{\widehat{t}(\pi)} m_0(t')dt' = V_-^F(\widehat{t}(\pi)) < V_-^F(t_-^F)$. Therefore for $\pi \in \left[\underline{\pi}, \frac{1}{\phi_1(t_-^F)+1}\right]$, F_- is optimal. Combining with Claim D.4.6, it is optimal for all $\pi \in \left(0, \frac{1}{\phi_1(t_-^F)+1}\right]$.

Similarly F_+ is optimal for all $\pi \in \left[\frac{1}{\phi_1(t_+^F)+1}, 1\right)$.

Similarly as in the previous paragraph, by Claim D.4.3, for $\pi \in \left[\frac{1}{\phi_1(t_-^F)+1}, \overline{\overline{\pi}}\right]$, V_-^F is strictly increasing in t over $\left\{t : \phi_1(t) \leq \frac{1-\pi}{\pi}\right\}$, so its maximum is achieved at $t = \widehat{t}(\pi)$. Hence for $\pi \in \left(\frac{1}{\phi_1(t_-^F)+1}, \frac{1}{\phi_1(t_+^F)+1}\right)$, $\arg\max_{\left\{t:\phi_1(t)\leq\frac{1-\pi}{\pi}\right\}} V_-^F(t) = \arg\max_{\left\{t:\phi_1(t)\geq\frac{1-\pi}{\pi}\right\}} V_+^F(t) = \widehat{t}(\pi)$, i.e. both optimal F_+ and F_- tests coincide to the simple T-F test.

D.5 Proofs for Section 5.2.2, Distortions

In Theorem 3.B, $\underline{t} = \min\{t_0^{fb}, t_1^{fb}, t^*\}$ and $\overline{t} = \max\{t_0^{fb}, t_1^{fb}, t^*\}$, where t^* is the unique signal threshold of the optimal test. Recall that there can be a maximum of one threshold, by regularity.

Define:

$$\frac{f(t|\omega, v)}{f(t|1,0)} = g_\omega(t|v,0) \;\forall\, \omega, t, v \tag{1}$$

Claim D.5.1. *Under MLRP, m_1 is increasing (m_0 is decreasing) whenever it is positive. Moreover, both m_1 and m_0 cross 0 only once.*

Proof.

$$m_1(t) = \left(\int_v v g(t|1,v) d\nu(v) \right) f(t|1,0)$$

Differentiating,

$$m_1'(t) = \left(\int_v v g_1(t|v,0) d\nu(v) \right) f'(t|1,0) + \left(\int_v v g'(t|1,v) d\nu(v) \right) f(t|1,0)$$

Note that by definition of the $g_1(t|v,0)$'s and MLRP, $g_1'(t|v,0) \gtreqless 0 \iff v \gtreqless 0$. Hence $\left(\int_v v g'(t|1,v) d\nu(v) \right) > 0$. Hence $m_1'(t) > 0$ if $\left(\int_v v g_1(t|v,0) d\nu(v) \right) > 0$, i.e. $m_1(t) > 0$. Therefore once $m_1(t)$ reaches zero from below it cannot turn back negative, which shows the "moreover" part for $m_1(t)$.

Identically as above, the results for $m_0(\cdot)$ follow.

□

Let $a_\omega^{fb} : T \to [0,1]$ denote the first-best allocation, $\omega \in \{0,1\}$. By Claim D.5.1, there exist $t_0^{fb} \in (0,1)$ and $t_1^{fb} \in (0,1)$ such that Let $a_0^{fb}(t) = 1(t \leq t_0^{fb}), a_1^{fb}(t) = 1(t \geq t_1^{fb})$. In particular, t_ω^{fb} solves $m_\omega(t_\omega^{fb}) = 0, \omega \in \{0,1\}$.

Claim D.5.2. $t_-^F > \min\{t_0^{fb}, t_1^{fb}\}$.

Proof. By Claim D.5.1, $m_0(t) > 0 \ \forall \ t \in [0, t_0^{fb}] \supseteq [0, \min\{t_0^{fb}, t_1^{fb}\}]$ and $m_1(t) < 0 \ \forall \ t \in [0, t_1^{fb}] \supseteq [0, \min\{t_0^{fb}, t_1^{fb}\}]$, i.e. $m_0(t) > 0 > m_1(t) \ \forall \ t \leq \min\{t_0^{fb}, t_1^{fb}\}$. Therefore each term in $V^{F\prime}{}_-(t) = -m_1(t) + \phi_1(t) m_0(t) + \phi_1'(t) \int_0^t m_0(t') dt'$ is positive for all $t \leq \min\{t_0^{fb}, t_1^{fb}\}$, $\therefore \ V^{F\prime}{}_-(t) > 0 \ \forall \ t \leq \min\{t_0^{fb}, t_1^{fb}\}$. By qausiconcavity of $V_L^F(t)$ the result follows. □

Let us assume $\pi < \underline{\pi}$ where $\underline{\pi}$ is as defined in Theorem 2.A. Therefore the optimal test is of type F_- with threshold t_-^F.

The case for $\pi > \overline{\pi}$ is reciprocal.

Claim D.5.3. *Suppose $\pi < \underline{\pi}$, where $\underline{\pi}$ as defined in Theorem 2.A. Then, the following table gives all possible distortions.*

The first column in each of the tables below captures intervals of types, and the table entries capture whether those types are better off (+), worse off (−) or face no distortions (0), compared to first-best.

	Cherry-picking, $t_1^{fb} > t_0^{fb}$	Lemon-dropping, $t_1^{fb} \leq t_0^{fb}$
$[0, t_0^{fb}]$	$-$	0
$[t_0^{fb}, \max\{t_1^{fb}, t_-^F\}]$	$+$	
$[\max\{t_1^{fb}, t_-^F\}, 1]$	0	$+$

Table 7: Distortions for various types in a competitive market

	Cherry-picking, $t_1^{fb} > t_0^{fb}$	Lemon-dropping, $t_1^{fb} \leq t_0^{fb}$
$[0, t_1^{fb}]$	$-$	0
$[t_1^{fb}, \min\{t_0^{fb}, t_-^F\}]$	$-$	
$[\min\{t_0^{fb}, t_-^F\}, \max\{t_0^{fb}, t_-^F\}]$	$+/- \iff t_-^F \gtrless t_0^{fb}$	
$[\max\{t_0^{fb}, t_-^F\}, 1]$	0	$+$

Table 8: Distortions for various types in an uncompetitive market

Proof. By Claim D.5.2, the above table captures all possible cases.

For $\pi < \underline{\pi}$, in a cherry-picking market, the optimal test with screening takes the form $\widehat{a}_0(t) = \left(\frac{\pi}{1-\pi}\right) \phi_1(t_-^F) \times 1(t \leq t_-^F), \widehat{a}_1(t) = 1(t \geq t_-^F)$. We prove the claims in the above tables for a cherry-picking market. The arguments for a lemon-dropping are almost identical, keeping in mind the fact that the optimal test with a threshold at t_-^F, in that case, is given by $\widehat{a}_0(t) = 1(t \leq t_-^F) + \left(1 - \frac{\pi}{1-\pi}\phi_1(t_-^F)\right) \times 1(t \geq t_-^F), \widehat{a}_1(t) = 1(t \geq t_-^F)$.

Let us first consider the case when $t_1^{fb} > t_0^{fb}$. Clearly, the types $t \leq t_0^{fb}$ are worse off under screening, as they are passed only in state 0 - as under the unconstrained solution - but with lower probability. Simiarly $t \in [t_0^{fb}, t_1^{fb}]$ are better off under screening, as they are never passed under first-best. If $t_-^F > t_1^{fb}$, $t \in [t_1^{fb}, t_-^F]$ are passed with probability 1 in state 1 and 0 in state 0 under the unconstrained solution. But given the test $(\widehat{a}_1, \widehat{a}_0)$ as described above, they can still choose this allocation, but choose not to, as the test is incentive compatible. Therefore they are better off with their allocation under screening. Hence $t \in [t_0^{fb}, \max\{t_1^{fb}, t_-^F\}]$ are better off. $t \geq \max\{t_1^{fb}, t_-^F\}$ obviously face no distortion - they are passed for sure, only in state 1, under both.

If $t_1^{fb} \leq t_0^{fb}$., naturally all $t \in [t_1^{fb}, t_0^{fb}]$ are worse off, as they are passed for sure under first-best. Naturally, in a cherry-picking market, so are all $t \leq t_1^{fb}$, as they are passed only in state 0 under both screening and first-best, but with lower probability under screening. If $t_-^F > t_0^{fb}$, analogously as argued in the previous paragraph for the competitive case, $t \in [t_0^{fb}, t_-^F]$, they are passed with probability one in state 0 - an option they have available under the screening test, but don't choose, due

to incentive compatibility. Therefore they are better off under screening. Types $t \geq \max\{t_0^{fb}, t_-^F\}$ again face no distortion, as in the previous case. □

D.6 Proofs for Section 5.3, Admission of uncertainty

For ease of defining objects to be introduced shortly, particularly for this subsection, we also define the notation π_ω, denoting the prior probability of state $\omega \in \{0,1\}$. Therefore in terms of our standard notation π for the prior, $\pi_1 = \pi, \pi_0 = 1-\pi$.

D.6.1 The structure of dual threshold tests

Consider a two-threshold IC mechanism with signal-thresholds $\underline{t}$ and $\overline{t} > \underline{t}$. Let the corresponding thresholds of normalized likelihood ratios be denoted by $\underline{s}_1, \overline{s}_1, \underline{s}_0, \overline{s}_0$ respectively. In particular, let:

$$\underline{s}_1 := \phi_1(\underline{t}), \overline{s}_1 := \phi_1(\overline{t}), \underline{s}_0 := \frac{1}{\underline{s}_1} = \phi_0(\underline{t}), \overline{s}_0 := \frac{1}{\overline{t}} = \phi_0(\overline{t}). \tag{1}$$

Let the belief thresholds corresponding to $\underline{t}$ and $\overline{t}$ be $\underline{p}$ and $\overline{p}$ respectively. Hence:

$$\begin{aligned} \frac{\underline{p}}{1-\underline{p}} &= \left(\frac{\pi}{1-\pi}\right)\phi_1(\underline{t}) = \frac{\underline{s}_1}{\widehat{s}_1} = \frac{\widehat{s}_0}{\underline{s}_0}, \\ \frac{\overline{p}}{1-\overline{p}} &= \left(\frac{\pi}{1-\pi}\right)\phi_1(\overline{t}) = \frac{\overline{s}_1}{\widehat{s}_1} = \frac{\widehat{s}_0}{\overline{s}_0}. \end{aligned} \tag{2}$$

Let the passing rates for the "middle region" of signals in our two threshold mechanism – $t \in [\underline{t}, \overline{t}]$ – be x and y in states 1 and 0 respectively. In other words, our two threshold mechanism is given by:

$$\boldsymbol{a}(t) := (\widehat{a}_1(t), \widehat{a}_0(t)) = \begin{cases} (0,1) \text{ for } t < \underline{t} \\ (x,y) \text{ for } t \in [\underline{t}, \overline{t}] \\ (1,0) \text{ for } t > \overline{t}. \end{cases} \tag{3}$$

D.6.2 Type 1 and Type 0 tests

Let $S_\omega = \phi_\omega([0,1]), \omega \in \{0,1\}$. Clearly, S_ω is an interval for each ω. Let s_ω denote a typical element of S_ω.

Fix a prior π. To maintain notational parity – as would be apparent shortly – we use the following notation:

$$\widehat{s}_1 := \frac{1-\pi}{\pi} \text{ and } \widehat{s}_0 := \frac{1}{\widehat{s}_1} = \frac{\pi}{1-\pi}$$

By Claim D.4.1 we can combine the pass and fail type tests with each undistorted state – 0 and 1 – into one, as follows. Let $\boldsymbol{\chi}^{\boldsymbol{\omega}} : S_\omega \to [0,1]^{2^{S_\omega}}, \omega \in \{0,1\}$ denote the

operator that maps each Normalized likelihood ratio (NLR, as defined in the main text, Section 5.3) to the *optimal* single threshold test (a pair of passing probabilities as a function of the reported signal) with that NLR as its threshold and undistorted state ω. Letting the first and second component denote $\widehat{a}_1$ and $\widehat{a}_0$ respectively:

$$\begin{aligned}\boldsymbol{\chi}^{\mathbf{1}}(s_1) &= \left(1(t \geq \phi_1^{-1}(s_1)), \left(\frac{s_1}{\widehat{s}_1}\right) 1(t \leq \phi_1^{-1}(s_1)) + \left(1 - \frac{s_1}{\widehat{s}_1}\right) v_0^+\right) \\ \boldsymbol{\chi}^{\mathbf{0}}(s_0) &= \left(\left(\frac{s_0}{\widehat{s}_0}\right) 1(t \geq \phi_0^{-1}(s_0)) + \left(1 - \frac{s_0}{\widehat{s}_0}\right) v_0^+, 1(t \leq \phi_0^{-1}(s_0))\right)\end{aligned} \tag{4}$$

We call χ^1 and χ^0 as Type 1 and Type 0 tests respectively.

Two points are worth emphasizing here. First, $\boldsymbol{\chi}^{\boldsymbol{\omega}}$, so defined, need not be feasible for all ω and $s_\omega \in S_\omega$, as we shall see shortly. Second, for any ω and $s_\omega \in S_\omega$, $\boldsymbol{\chi}^{\boldsymbol{\omega}}(s_\omega)$ so defined is *optimal* – not globally, but – under the restrictions of the fixed threshold s_ω and keeping ω the undistorted state.

We can also define the principal's value from each type of test as a function of its NLR-threshold as follows:

$$\begin{aligned}V_1(s_1) &= \pi\left(s_1\left(M_0(\phi_1^{-1}(s_1)) - v_0^+\right) + M_1(\phi_1^{-1}(s_1)) + \widehat{s}_1 v_0^+\right) \\ V_0(s_0) &= (1-\pi)\left(s_0\left(M_1(\phi_1^{-1}(s_1)) - v_0^+\right) + M_0(\phi_0^{-1}(s_0)) + \widehat{s}_0 v_0^+\right)\end{aligned} \tag{5}$$

D.6.3 The proof

We call a test $\boldsymbol{\chi}^{\boldsymbol{\omega}}(\underline{s}_\omega)$ IC if it, as defined by (4), satisfies the IC constraints (IC), and feasible if it satisfies (Feas). Note that all IC tests need not be feasible.

Claim D.6.1. *For any $\underline{s}_\omega < \overline{s}_\omega$ and π_ω such that $\frac{1-\pi_\omega}{\pi_\omega} \in (\underline{s}_\omega, \overline{s}_\omega)$, the following is the same mechanism for both $\omega \in \{0,1\}$, and it is feasible and IC:*

$$\left(\frac{\overline{s}_\omega - \widehat{s}_\omega}{\overline{s}_\omega - \underline{s}_\omega}\right) \boldsymbol{\chi}^{\boldsymbol{\omega}}(\underline{s}_\omega) + \left(\frac{\widehat{s}_\omega - \underline{s}_\omega}{\overline{s}_\omega - \underline{s}_\omega}\right) \boldsymbol{\chi}^{\boldsymbol{\omega}}(\overline{s}_\omega)$$

Proof. We show this by showing that any dual threshold feasible IC test can be expressed as a convex combination of two single threshold IC tests – as shown above – exactly one of which is feasible.

The threshold types of the dual threshold test – with beliefs $\underline{p}$ and $\overline{p}$ – must be indifferent between reporting their "left" and "right" messages. This gives us:

$$\begin{aligned}y(1-\underline{p}) + x\underline{p} = 1 - \underline{p} &\implies x = \left(\frac{1-\underline{p}}{\underline{p}}\right)(1-y), \\ y(1-\overline{p}) + x\overline{p} = \overline{p} &\implies (1-x) = \left(\frac{1-\overline{p}}{\overline{p}}\right) y.\end{aligned} \tag{6}$$

Solving the equation system (6) simultaneously we have:

$$x = \frac{\overline{s}_1 - \widehat{s}_1}{\overline{s}_1 - \underline{s}_1},\ y = \frac{\underline{s}_0 - \widehat{s}_0}{\underline{s}_0 - \overline{s}_0}. \tag{7}$$

The mechanism $\boldsymbol{a}$ can be written as:

$$\begin{aligned}
\widehat{a}_1(t) &= x1(t \geq \underline{t}) + (1-x)1(t \geq \overline{t}), \\
\widehat{a}_0(t) &= (1-y)1(t \leq \underline{t}) + y1(t \leq \overline{t}) \\
&= x\left(\frac{\underline{p}}{1-\underline{p}}\right)1(t \leq \underline{t}) + (1-x)\left(\frac{\overline{p}}{1-\overline{p}}\right)1(t \leq \overline{t}).
\end{aligned} \tag{8}$$

The last line comes from using (6).

Note that:

$$x\left(1 - \frac{\underline{p}}{1-\underline{p}}\right) + (1-x)\left(1 - \frac{\overline{p}}{1-\overline{p}}\right) = 0. \tag{Vanishing}$$

The above comes from the expressions of x and y in (7).

Combining (8) and (Vanishing), $\boldsymbol{a}$ can be written as:

$$\boldsymbol{a} = x\boldsymbol{\chi}^{\mathbf{1}}(\underline{t}) + (1-x)\boldsymbol{\chi}^{\mathbf{1}}(\overline{t}) \tag{9}$$

Because of (Vanishing), the $\left(1 - \frac{\mu(t_0)}{1-\mu(t_0)}\right)1(\omega = 0, V_0 \geq 0)$ term in $\boldsymbol{\chi}^{\mathbf{0}}$ plays no role, even when $V_0 \geq 0$.

Similarly as (Vanishing), we also have, from the expression for y in (7), $(1-y)\left(1 - \frac{1-\underline{p}}{\underline{p}}\right) + y\left(1 - \frac{1-\overline{p}}{\overline{p}}\right) = 0$.. Using this we also have:

$$\boldsymbol{a} = (1-y)\boldsymbol{\chi}^{\mathbf{0}}(\underline{t}) + y\boldsymbol{\chi}^{\mathbf{0}}(\overline{t})$$

□

Let $V^D : T \times T \to \mathbb{R}$ denote the principal's value from a dual threshold test as a function of its two signal-thresholds.

Claim D.6.2. *The principal's value from a two-threshold IC test with signal thresholds $\underline{t} < \overline{t}$ is given by:*

$$\begin{aligned}
V^D(\underline{t}, \overline{t}) &= xV_1(\phi_1(\underline{t})) + (1-x)V_1(\phi_1(\overline{t})) \\
&= yV_0(\phi_0(\underline{t})) + (1-y)V_0(\phi_0(\overline{t})).
\end{aligned} \tag{10}$$

where x and y are given by (7).

Proof. Follows directly from Claim D.6.1. □

Claim D.6.3. *If $\pi \leq \frac{1}{\phi_1(t_1^*)+1}$ or $\pi \geq \frac{1}{\phi_1(t_0^*)+1}$, the optimal mechanism must have a single threshold.*

Proof. By Claim D.6.2, the value from any dual threshold IC test is equal to some convex combination of two single threshold tests of each type. Therefore it is weakly lower than the maximum value of each of the two types of tests, i.e. for any dual threshold test with signal thresholds $\underline{t} < \overline{t}$, by (10),

$$V^D(\underline{t}, \overline{t}) \leq \min\{V_1(\phi_1(t_1^*)), V_0(\phi_0(t_0^*))\}$$

If $\pi \leq \frac{1}{\phi_1(t_1^*)+1}$, the type 1 test with signal threshold t_1^* is feasible and if $\pi \geq \frac{1}{\phi_1(t_0^*)+1}$ the type 0 test with signal threshold t_0^* is feasible. Using one of these the principal can do weakly better than any dual threshold test. Hence the claim follows. □

The following two corollaries follow.

Corollary D.6.1. *If $t_0^* \geq t_1^*$, the optimal mechanism must have a single threshold for all priors.*

Corollary D.6.2. *If $t_0^* < t_1^*$, the optimal mechanism must have a single threshold for all priors $\pi \notin \left(\frac{1}{\phi_1(t_1^*)+1}, \frac{1}{\phi_1(t_0^*)+1}\right)$.*

By the above corollaries, going forward we assume $t_0^* < t_1^*$ and consider the case $\pi \in \left(\frac{1}{\phi_1(t_1^*)+1}, \frac{1}{\phi_1(t_0^*)+1}\right)$.

Next we show that the principal's value is given by the maximium of her values from the monotone concave envelope of her value functions from the two types of tests – 1 and 0. The monotone concave envelope of a function is the lowest monotone and concave function which lies weakly above it everywhere. Specifically, the monotone concave envelope of $v_\omega, \omega \in \{0,1\}, \widehat{\widehat{v}}_\omega : \phi_\omega([0,1]) \to \mathbb{R}$, is defined as $\widehat{\widehat{v}}_\omega : s_\omega \mapsto \max_{\widetilde{s}_\omega \leq s_\omega} \widehat{v}_\omega(\widetilde{s}_\omega)$.

Claim D.6.4. *For any prior π, the principal's maximized value is given by:*

$$\max\left\{\widehat{\widehat{v}}_1\left(\frac{1-\pi}{\pi}\right), \widehat{\widehat{v}}_0\left(\frac{\pi}{1-\pi}\right)\right\} \qquad \text{(Monotone Concave Envelope)}$$

Proof. Note that if the best single-threshold test of either type is feasible then it is optimal.

We first find the principal's maximized value from each type of test - 1 and 0 - of a given threshold. We provide the proof for type 1 tests. The proof for type 0 tests is identical.

Clearly, if the prior is low enough, i.e. if $\phi_1(t_1^*) \leq \widehat{s}_1$, t_1^* is feasible as the threshold of a type 1 test. Therefore for $\phi_1(t_1^*) \leq \widehat{s}_1$, the principal-optimal single threshold test of type 1 is one with signal threshold t_1^*.

For $\phi_1(t_1^*) > \widehat{s}_1$, by Claim D.6.2, the maximized value from type 1 tests is given by:

$$\max_{\underline{s}_1,\overline{s}_1,\widehat{s}_1\in[\underline{s}_1,\overline{s}_1]} \left(\frac{\overline{s}_1-\widehat{s}_1}{\overline{s}_1-\underline{s}_1}\right) V_1(\underline{s}_1) + \left(\frac{\widehat{s}_1-\underline{s}_1}{\overline{s}_1-\underline{s}_1}\right) V_1(\overline{s}_1)$$

This is clearly the expression for $\widehat{v}_1(\widehat{s}_1)$ – the concave envelope of V_1 evaluated at $\widehat{s}_1 = \frac{1-\pi}{\pi}$.

Combining both cases – $\phi_1(t_1^*) \leq \widehat{s}_1$ and $\phi_1(t_1^*) > \widehat{s}_1$ – the principal's maximized value from a test of type 1 as a function of $\widehat{s}_1$ is given by:

$$V_1^*(\widehat{s}_1) := \begin{cases} V_1(\phi_1(t_1^*)), \phi_1(t_1^*) \leq \widehat{s}_1, \\ \widehat{v}_1(\widehat{s}_1), \phi_1(t_1^*) > \widehat{s}_1. \end{cases}$$

Note that the above is the expression for the monotone concave envelope of $V_1, \widehat{\widehat{v}}_1 : \phi_1([0,1]) \to \mathbb{R}$, defined as $\widehat{\widehat{v}}_1 : s_1 \mapsto \max_{\widetilde{s}_1 \leq s_1} \widehat{v}_1(\widetilde{s}_1)$.

Similarly as above, the principal's maximized value from a test of type 0 as a function of $\widehat{s}_0$ is also given by the monotone concave envelope of $\widehat{v}_0$:

$$(\widehat{s}_0) := \begin{cases} V_0(\phi_0(t_0^*)), \phi_0(t_0^*) \leq \widehat{s}_0, \\ \widehat{v}_0(\widehat{s}_0), \phi_0(t_0^*) > \widehat{s}_0. \end{cases}$$

Since the principal can choose any of the types of tests – 1 or 0 – her maximized value is given by (Monotone Concave Envelope).

□

Completing the proof. The equivalence of the first and second bullet points is clear from (10) and the fact that if $t \in (t_0^*, t_1^*)$, when $\phi_1(t) = \frac{1-\pi}{\pi}$, $V_1(\phi_1(t)) = V_0(\phi_0(t)) =$ the principal's value from the simple True-False test, which, for this prior, is the bang-bang test with belief-threshold $= \frac{1}{2}$. Hence, $\widehat{v}_1(\phi_1(t)) > V_1(\phi_1(t)) \iff \widehat{v}_0(\phi_0(t)) > V_0(\phi_0(t))$. Consequently, by Claim D.6.4 and Corollary D.6.2 the result follows.

D.7 Proofs for Section 6: Endogenous topic selection

Proposition 13. *The principal's maximized value is concave in the prior.*

Let us denote the principal's maximized value for prior π as $V(\pi)$. Fix $s_1 \in (0,1)$ and $\pi_1, \pi_2 \in [0,1]$. We have to show that,

$$V(s_1\pi_1 + (1-s_1)\pi_2) \geq s_1 V(\pi_1) + (1-s_1)V(\pi_2) \tag{1}$$

Let $\pi := s_1\pi_1 + (1-s_1)\pi_2$.

Consider the problem where the principal's prior is π and she has access to a fixed binary experiment – one that produces two posteriors, π_1 and π_2, with probability s_1 and $s_2 = 1 - s_1$ respectively. She can commit to implement the mechanism $(\widehat{a}_1^i, \widehat{a}_0^i) : T \to [0,1]$ if her posterior is $\pi_i, i \in \{1,2\}$. Her payoff under this mechanism is:

$$
\begin{aligned}
&= \sum_{i\in\{1,2\}} s_i \left(\pi_i \int_0^1 m_1(t)\widehat{a}_1^i(t)dt + (1-\pi_i) \int_0^1 m_0(t)\widehat{a}_0^i(t)dt \right) \\
&= \pi \int_0^1 m_1(t)\widetilde{a}_1(t)dt + (1-\pi) \int_0^1 m_0(t)\widetilde{a}_0(t)dt
\end{aligned}
$$

where $\widetilde{a}_1(t) = \frac{\sum_{i\in\{1,2\}} s_i\pi_i\widehat{a}_1^i(t)}{\pi}$, and $\widetilde{a}_0(t) = \frac{\sum_{i\in\{1,2\}} s_i(1-\pi_i)\widehat{a}_0^i(t)}{1-\pi}$.

Hence, she solves the following problem:

$$
\max_{\widehat{a}_1^1,\widehat{a}_0^1,\widehat{a}_1^2,\widehat{a}_0^2\in[0,1]^4} \quad \pi \int_0^1 m_1(t)\widetilde{a}_1(t)dt + (1-\pi) \int_0^1 m_0(t)\widetilde{a}_0(t)dt \tag{2}
$$

$$
\text{s.t. } \widehat{a}_1^1, \widehat{a}_0^1, \widehat{a}_1^2, \widehat{a}_0^2 \in [0,1], \text{and } \forall\, t, t' \in T,
$$

$$
\sum_{i\in\{1,2\}} s_i \left(\mu(t;\pi_i)\widehat{a}_1^i(t) + (1-\mu(t;\pi_i))\widehat{a}_0^i(t)\right) \geq \sum_{i\in\{1,2\}} s_i \left(\mu(t;\pi_i)\widehat{a}_1^i(t') + (1-\mu(t;\pi_i))\widehat{a}_0^i(t')\right), \tag{IC}
$$

where $\widetilde{a}_1(t) = \dfrac{\sum_{i\in\{1,2\}} s_i\pi_i\widehat{a}_1^i(t)}{\pi}$, and $\widetilde{a}_0(t) = \dfrac{\sum_{i\in\{1,2\}} s_i(1-\pi_i)\widehat{a}_0^i(t)}{1-\pi}$.

In the above problem, note that if we restrict the principal to use $(\widehat{a}_1^1, \widehat{a}_0^1) = (\widehat{a}_1^2, \widehat{a}_0^2) = (\widetilde{a}_1, \widetilde{a}_0)$, we get back the (signal-based version of the) original problem, (1). Hence, since (1) allows the principal to choose from a subset of the mechanisms that (2) allows, the value of (1) is weakly less than that of (2). But by the revelation principle, the value of (2) is weakly less than that of (1). Hence their values are equal.

Now consider the alternative case where the principal is restricted to revealing her signal to the agent. In this case she solves:

$$
\max_{\widehat{a}_1^1,\widehat{a}_0^1,\widehat{a}_1^2,\widehat{a}_0^2\in[0,1]^4} \quad \pi \int_0^1 m_1(t)\widetilde{a}_1(t)dt + (1-\pi) \int_0^1 m_0(t)\widetilde{a}_0(t)dt \tag{3}
$$

$$
\text{s.t. } \widehat{a}_1^1, \widehat{a}_0^1, \widehat{a}_1^2, \widehat{a}_0^2 \in [0,1], \text{and } \forall\, t, t' \in T, i \in \{1,2\},
$$

$$
\mu(t;\pi_i)\widehat{a}_1^i(t) + (1-\mu(t;\pi_i))\widehat{a}_0^i(t) \geq \mu(t;\pi_i)\widehat{a}_1^i(t') + (1-\mu(t;\pi_i))\widehat{a}_0^i(t'). \tag{IC-restricted}
$$

Clearly, any mechanism which satisfies (IC-restricted) satisfies (IC). Let the value of (3) be denoted by V_{restr}. Then, we must have $V_{restr} \leq V(\pi)$.

Now consider a third case where the principal must reveal her posterior to the agent, like the previous case, but in addition, cannot pre-commit to mechanisms as a function of her posteriors. Hence, she must choose a mechanism *after* observing her posterior. We call the principal who observes posterior π_i, the principal's i-th *interim self*.

V_{restr} is the principal's ex-ante value when the mechanism $(\widehat{a}_1^1, \widehat{a}_0^1, \widehat{a}_1^2, \widehat{a}_0^2)$ is chosen by her ex-ante self subject to (IC-restricted). Hence her ex-ante value from it must be weakly higher than her value, if, each of the mechanisms $(\widehat{a}_1^1, \widehat{a}_0^1)$ and $(\widehat{a}_1^2, \widehat{a}_0^2)$ were chosen by her corresponding interim self, with posteriors π_1 and π_2 respectively, because they would be facing the same IC constraints, (IC-restricted).

The i-th interim self of the principal solves:

$$\begin{aligned}\max_{\widehat{a}_1^i, \widehat{a}_0^i \in [0,1]^2} \quad & \pi_i \int_0^1 m_1(t)\widehat{a}_1^i(t)dt + (1-\pi_i)\int_0^1 m_0(t)\widehat{a}_0^i(t)dt \\ & \text{s.t. } \widehat{a}_1^i, \widehat{a}_0^i \in [0,1], \text{and } \forall\, t, t' \in T, \\ \mu(t;\pi_i)\widehat{a}_1^i(t) + (1-\mu(t;\pi_i))\widehat{a}_0^i(t) &\geq \mu(t;\pi_i)\widehat{a}_1^i(t') + (1-\mu(t;\pi_i))\widehat{a}_0^i(t').\end{aligned}$$

The principal's *ex-ante* value from each interim self choosing *its* optimal mechanism is, therefore:

$$\begin{aligned}\max_{\widehat{a}_1^1, \widehat{a}_0^1, \widehat{a}_1^2, \widehat{a}_0^2 \in [0,1]^4} \quad & \sum_{i\in\{1,2\}} s_i \left(\pi_i \int_0^1 m_1(t)\widehat{a}_1^i(t)dt + (1-\pi_i)\int_0^1 m_0(t)\widehat{a}_0^i dt \right) \\ & \text{s.t. } \widehat{a}_1^1, \widehat{a}_0^1, \widehat{a}_1^2, \widehat{a}_0^2 \in [0,1], \text{and } \forall\, t, t' \in T, i \in \{1,2\}, \\ \mu(t;\pi_i)\widehat{a}_1^i(t) + (1-\mu(t;\pi_i))\widehat{a}_0^i(t) &\geq \mu(t;\pi_i)\widehat{a}_1^i(t') + (1-\mu(t;\pi_i))\widehat{a}_0^i(t').\end{aligned} \tag{4}$$

Clearly, the feasible set of mechanisms for (3) and (4) are the same, but the objectives are (potentially) different. The principal's ex-ante self therefore prefers the one where the objective is the principal's ex-ante value, i.e. (3). That is,

$$V_{restr} \geq \sum_{i\in\{1,2\}} s_i V(\pi_i)$$

Proof of Theorem 4. We prove the claims for the cherry-picking case. The proofs for the lemon-dropping case are similar.

First bullet point. Clearly, the principal's value is linearly increasing (decreasing) in π for $(\pi \in [0, \underline{\pi}])$ $(\pi \in [\overline{\pi}, 1])$. Therefore her optimal π's must lie in $[\underline{\pi}, \overline{\pi}]$. Below we show that they lie in $(\underline{\pi}, \overline{\pi})$.

Note that the principal's maximized value for all $\pi \in (\underline{\pi}, \overline{\pi})$, is given by $\mathbb{V}(\pi) = V_-^F(\widehat{t}(\pi)) = V_+^F(\widehat{t}(\pi))$. Denoting $V_-^F(t) = \pi V_{-,0}^F(t)$ where $V_{-,0}^F(t) := \phi_1(t)\int_0^t m_0(t')dt' + \int_0^t m_1(t')dt'$, we have, for $\pi \in [\underline{\pi}, \overline{\pi}]$:

$\mathbb{V}'(\pi) = \pi V^{F\prime}_{-,0}(\widehat{t}(\pi))\underline{t}'(\pi) + V_{-,0}^F(\widehat{t}(\pi))$

At $\pi = \underline{\pi}, \underline{t}(\underline{\pi}) = t_-^F$, $\therefore\ V^{F\prime}_{-,0}(\underline{t}(\underline{\pi})) = 0, V_{-,0}^F(\underline{t}(\underline{\pi})) > 0$ since we assumed the solution is not constant. Hence $\mathbb{V}(\pi)$ is strictly increasing at $\pi = \underline{\pi}$. Similarly, using the formulation $\mathbb{V}(\pi) = V_+^F(\widehat{t}(\pi))$ it can be shown that it is strictly decreasing at $\pi = \overline{\pi}$. The claim follows.

Second bullet point. Follows directly from the first and Theorem 2.A for the regular case.

For the general case, note that the principal's value from a two-threshold test with signal thresholds $\underline{t}$ and $\overline{t}$ are given by:

$$\left(\frac{\pi\phi_1(\overline{t}) - (1-\pi)}{\phi_1(\overline{t}) - \phi_1(\underline{t})}\right) V_-^F(\underline{t}) + \left(\frac{(1-\pi) - \pi\phi_1(\underline{t})}{\phi_1(\overline{t}) - \phi_1(\underline{t})}\right) V_-^F(\overline{t})$$

The above is linear in π. Hence the optimum is achieved at one of the extremes, making the optimal test of the simple T-F type.

Third bullet point. Total passing probability under the simple T-F mechanism:

$$\begin{aligned}
&= \pi \int_{\mu(t;\pi)\geq\frac{1}{2}} \overline{f}_1(t)dt + (1-\pi)\int_{\mu(t;\pi)\leq\frac{1}{2}} \overline{f}_0(t)dt \\
&= \int_{\pi\overline{f}_1(t)\geq(1-\pi)\overline{f}_0(t)} \pi\overline{f}_1(t)dt + \int_{\pi\overline{f}_1(t)\leq(1-\pi)\overline{f}_0(t)} (1-\pi)\overline{f}_0(t)dt \\
&= \int_0^1 \max\{\pi\overline{f}_1(t), (1-\pi)\overline{f}_0(t)\}dt \\
&\geq \max\left\{\pi\int_0^1 \overline{f}_1(t), (1-\pi)\int_0^1 \overline{f}_0(t)\right\}dt \\
&= \max\{\pi, 1-\pi\} \\
&\geq \frac{1}{2}.
\end{aligned}$$

Fourth bullet point. Follows directly from symmetry and Proposition 13.

D.8 Proofs for Section 6, Other equilibria of the informed principal game

Proof of Theorem 5. Define $M_0, M_1 : [0,1] \to \mathbb{R}$ as $M_0(t) := \int_0^t m_0(t')dt'$ and $M_1(t) := \int_t^1 m_1(t')dt'$ for all t. Further, let $\gamma := \left(\frac{\pi}{1-\pi}\right)\left(\frac{1-\pi_P}{\pi_P}\right)$.

Note that Claim D.4.1 goes through even in this case. Hence, the expression for the principal's value from a test with a belief threshold below one half – the analog of (1) – is given by:

$$\begin{aligned} V_-(t) &:= \pi_P \int_t^1 m_1(t')dt' + (1-\pi_P)\left[\left(\frac{\pi_A}{1-\pi_A}\right)\left(\frac{\mu(t;\pi)}{1-\mu(t;\pi)}\right)\int_0^t m_0(t')dt' \right. \\ &\quad \left. + \left(1 - \left(\frac{\pi_A}{1-\pi_A}\right)\left(\frac{\mu(t;\pi)}{1-\mu(t;\pi)}\right)\right)V_0^+\right] \\ &= \pi_P\left(M_1(t) + \gamma\phi_1(t)(M_0(t) - V_0^+)\right) + (1-\pi_P)V_0^+ \end{aligned} \tag{1}$$

where $V_0^+ = \max\{0, V_0\}$.

Differentiating,

$$V_-'(t) = \pi_P\left(-m_1(t) + \gamma\left(\phi_1(t)m_0(t) + \phi_1'(t)(M_0(t) - V_0^+)\right)\right) \tag{2}$$

The first-order condition, analogous to (3), is given by:

$$\underbrace{\frac{\phi_1'(t)(M_0(t) - V_0^+) + \phi_1(t)m_0(t)}{m_1(t)}}_{=:L(t)} = \frac{1}{\gamma} \tag{3}$$

Let us call the LHS of (3) $L(t)$, as shown. Let us also denote the numerator as a function of t by $N_-(t)$.

Recall that t_ω^{fb} was defined as $m_\omega(t_\omega^{fb}) = 0, \omega \in \{0,1\}$.

Claim D.8.1. *For all $\gamma \in [0,\infty)$, (3) has exactly one solution.*

Proof. First we show that for each $\gamma \in [0,\infty)$, (3) has at most one solution.

Analogously as (4), if t is such that ${V^F}'_-(t) = 0$, ${V^F}''_-(t) < 0$ by strong regularity, regardless of γ. Hence $V_-(\cdot)$ is single-peaked, as in the baseline case. Hence (3) has at most one solution for each $\gamma \in (0,\infty)$.

Now we show that for any $\gamma \in [0,\infty)$ there exists $t \in (0,1)$ such that (3) is satisfied.

First, consider the case when $N_-(t_1^{fb}) = 0$, i.e. the numerator and denominator of $L(t)$ vanish at the same point. In this case, by (2), $V_-'(t_1^{fb}) = 0$ for all γ and we

are done. Hence for the rest of the proof we assume $N_-(t_1^{fb}) \neq 0$. Hence $\lim_{t \to t_1^{fb}}$ does not exist.

First, consider the case when $V_0 \leq 0$. Clearly, for t sufficiently close to 0, $m_1(t) < 0$ and similarly $m_0(t') > 0$ for all $t' \leq t$. Hence $M_0(t) > 0$. Hence from (2), ${V^F}'_-(t) > 0$. Similarly, at $t = 1$, $M_0(t) = V_0 \leq 0$, $m_0(t) < 0$, $m_1(t) > 0$. By continuity of $\phi_1'(t)$ and each of the other terms in the expression for ${V^F}'_-(t)$, ${V^F}'_-(t)$ is continuous and therefore attains the value 0 for some $t \in (0, 1)$, for each γ.

Now consider the case when $V_0 > 0$. At $t = 1, (M_0(t) - V_0) = 0, m_0(t) < 0, m_1(t) > 0$, hence $V'_-(1) < 0$. At $t = t_0^{fb}$, $M_0(t)$ attains its maximum. Hence $(M_0(t) - V_0) > 0$. By definition of t_0^{fb}, $m_0(t_0^{fb}) = 0$. Hence $N_-(t_0^{fb}) > 0$.

First consider the case when $m_1(t_0^{fb}) > 0$. In this case $L(t_0^{fb}) > 0$. This case also implies $t_1^{fb} < t_0^{fb}$, hence $m_1(t) > 0$ for all $t \geq t_0^{fb}$. Therefore by continuity of $L(t)$ in $[t_0^{fb}, 1]$ we are done.

Now consider the case when $m_1(t_0^{fb}) < 0$. In this case $t_1^{fb} > t_0^{fb}$ and $L(t_0^{fb}) < 0$. By assumption, $N_-(t_1^{fb}) \neq 0$. If $N_-(t_1^{fb}) > 0$, $\lim_{t \downarrow t_1^{fb}} = \infty$, hence $L(t_+) = 0$ for some $t_+ \in (t_1^{fb}, 1)$ and we are done. If $N_-(t_1^{fb}) < 0$, $\lim_{t \uparrow t_1^{fb}} = \infty$, hence $L(t_-) = 0$ for some $t_- \in (t_0^{fb}, t_1^{fb})$ and we are done.

□

By Claim D.8.1, for every $\gamma \in (0, \infty)$ there exists exactly one t such that (3) holds. Hence we can define the function $t_-^*(\gamma)$ which solves (3), as a function of γ.

Claim D.8.2. *Either $t_-^*(\gamma) = t_1^{fb}$ for all $\gamma \in [0, \infty)$, or exactly one of the following statements holds:*

- $\lim_{t \uparrow t_1^{fb}} L(t) = \infty, \lim_{t \downarrow t_1^{fb}} L(t) = -\infty$, *there exists $t_- \in (0, t_1^{fb})$ such that $L(t)$ is strictly positive and monotonically increasing for $t \in (t_-, t_1^{fb})$ and $L(t) < 0$ for all $t > t_1^{fb}$.*
- *There exists $t_+ \in (t_1^{fb}, 1)$ such that $L(t_+) = 0$, $\lim_{t \downarrow t_1^{fb}} L(t) = \infty, \lim_{t \uparrow t_1^{fb}} L(t) = -\infty$, $L(t)$ is strictly positive and monotonically decreasing for $t \in (t_1^{fb}, t_+)$ and $L(t) < 0$ for all $t < t_1^{fb}$.*

Proof. If $N_-(t_1^{fb}) = 0$, the first case arises – $t_-^*(\gamma) = t_1^{fb}$ for all $\gamma \in [0, \infty)$ – and we are done. Hence for the rest of the proof we assume $N_-(t_1^{fb}) \neq 0$.

Since t_1^{fb} is the only point of discontinuity of $L(t)$, this means we cannot have $\lim_{t \uparrow t_1^{fb}} L(t) = \lim_{t \downarrow t_1^{fb}} L(t) = -\infty$, because in that case, $L(t)$ cannot take all values in $[0, \infty)$ for $t \in [0, t_1^{fb}) \cup (t_1^{fb}, 1]$, in violation of Claim D.8.1. Hence either $\lim_{t \uparrow t_1^{fb}} L(t) = \infty$ or $\lim_{t \downarrow t_1^{fb}} L(t) = \infty$ or both.

Suppose $\lim_{t\uparrow t_1^{fb}} L(t) = \lim_{t\downarrow t_1^{fb}} L(t) = \infty$. Hence for a large enough $L_0 > 0$, there exist $t < t_1^{fb}$ and $t' > t_1^{fb}$ such that $L(t) = L(t') = L_0$, contradicting Claim D.8.1.

Therefore, either $\lim_{t\uparrow t_1^{fb}} L(t) = \infty$ or $\lim_{t\downarrow t_1^{fb}} L(t) = \infty$ but not both. Since, $N_-(t_1^{fb}) \neq 0$ by assumption, either $\lim_{t\uparrow t_1^{fb}} L(t) = \infty$ and $\lim_{t\downarrow t_1^{fb}} L(t) = -\infty$ or $\lim_{t\uparrow t_1^{fb}} L(t) = -\infty$ and $\lim_{t\downarrow t_1^{fb}} L(t) = \infty$. Strict monotonicity of $L(t)$ on opposite sides of t_1^{fb} are ensured by Claim D.8.1, due to the fact that L takes all non-negative values exactly once. □

Claim D.8.3. *Case II from Claim D.8.1 arises if and only if* $N_-(t_1^{fb}) > 0$.

Proof. Obvious given Claim D.8.2. □

In the rest of the proof, we show that Case II from Claim D.8.2 must arise under our assumptions.

Claim D.8.4. *If* $\max_{t\in[0,1]} \min\{m_1(t), m_0(t)\} \geq 0$, *Case II from Claim D.8.2 must arise.*

Proof. First consider the case $V_0 \leq 0$.

If $\max_{t\in[0,1]} \min\{m_1(t), m_0(t)\} \geq 0$ – and therefore $M_0(t) > 0$ – for all $t \leq t_1^{fb}$. Therefore the numerator of $L(t)$ is strictly positive at $t = t_1^{fb}$. Hence, by continuity, there exists $\epsilon > 0$ such that it is strictly positive for all $t \in (t_1^{fb}, t_1^{fb} + \epsilon)$. $m_1(t) < 0$ for $t < t_1^{fb}$ and $m_1(t) > 0$ for $t > t_1^{fb}$. Hence $\lim_{t\uparrow t_1^{fb}} L(t) = -\infty$ and $\lim_{t\downarrow t_1^{fb}} L(t) = \infty$, i.e. Case II from Claim D.8.2 must arise.

Next, suppose $V_0 > 0$.

If $\max_{t\in[0,1]} \min\{m_1(t), m_0(t)\} \geq 0$, $m_1(t_0^{fb}) > 0$ and $t_1^{fb} < t_0^{fb}$. As argued for Claim D.8.1, in this case $L(t_0^{fb}) > 0$. By the fact that $t_1^{fb} < t_0^{fb}$, this violates the last part of Case I of Claim D.8.2. Hence case II must arise. □

Completing the proof. When Case II arises, $t_-^*(\gamma)$ is increasing in γ, and therefore in π_A, for a fixed π_P. As we showed earlier, $\underline{t}(\pi_A)$ is decreasing in π_A. Hence the arguments in the "completing the proof" part of the proof of Theorem 2.A go through and the expression for $\underline{\pi}(\pi_P)$ is derived accordingly. Similar arguments as above show that the "upper" signal threshold, analogous to t_+^P and t_+^F – let us call it $t_+^*(\gamma)$ – is also increasing in γ. This gives us the existence and expression for $\overline{\pi}(\pi_P)$.

Proof of Prop 7. We have to show that under the optimal mechanism for $\pi_P = \pi$, the principal's value *in each state* is at least as much as the maximum value she can obtain if she reveals the state. In the latter case, there is no screening, so her maximum value is $= \max\{0, V_0\}$. Hence we have to show the following:

$$\min\left\{\int m_1(t)\widehat{a}_1(t)dt, \int m_0(t)\widehat{a}_0(t)dt\right\} \geq \max\{0, V_0\} \;\forall\; \pi_P \in [0,1]. \tag{4}$$

where $(\widehat{a}_1, \widehat{a}_0)$ is the optimal mechanism with $\pi_P = \pi$.

We consider two cases - when the optimal mechanism is constant, when it has a single threshold.

Case I: There exists an optimal mechanism which is constant. In this case the maximized value is equal to $\max\{0, V_0\}$, which is equal to the principal's maximum value if she discloses the state. Hence she cannot do strictly better by disclosing the state, in any of the states. Hence such a mechanism is a core mechanism.

For the rest of the cases we assume there does not exist a constant mechanism which is optimal, i.e. the principal's optimal mechanism for $\pi_P = \pi$ gives her strictly greater value than any constant mechanism.

Case II: The optimal mechanism has a single threshold.

We show this for the case when the optimal test is of type F_-. The proofs for the other three cases are similar. Consider a F_- test with a belief threshold $\underline{p}$. We must have $\underline{p} \leq \frac{1}{2}$. Letting $(\widetilde{a}_1, \widetilde{a}_0)$ denote the corresponding mechanism in terms of beliefs, and letting $q = \widetilde{a}_1 - \widetilde{a}_0$, we have:

$$\begin{aligned} q(p) &= \left(-\frac{\underline{p}}{1-\underline{p}}\right) 1(p \leq \underline{p}) + 1(p > \underline{p}) \\ &= \left(\frac{\underline{p}}{1-\underline{p}}\right) \times \underbrace{(-1(p \leq \underline{p}) + 1(p \geq \underline{p}))}_{q_1} + \left(1 - \frac{\underline{p}}{1-\underline{p}}\right) \times \underbrace{1}_{q_2} \end{aligned}$$

If F_- is optimal we must have $V_0 \leq 0$. Hence the optimal q solves either (4) or (10). In both cases the objective function is quasiconvex in q. Hence the optimized value is dominated by the maximum of its values evaluated at q_1 and q_2, where q_1 and q_2 are as shown above. As we see above, q is a convex combination of q_1 and q_2. Clearly, q_2 corresponds to a feasible, IC mechanism which is constant. Hence by our assumption, the optimized objective must be strictly greater than under that mechanism. Hence it must be weakly dominated by the objective under the (non-IC) mechanism corresponding to q_1 - the bang-bang mechanism with a threshold at $\underline{p}$. Letting $\underline{t} \in [0,1]$ be such that $\mu(\underline{t}; \pi) = \underline{p}$, that is equivalent to:

$$\begin{aligned} &\pi \int_{\underline{t}}^{1} m_1(t)dt + (1-\pi)\left(\frac{\underline{p}}{1-\underline{p}}\right) \int_{0}^{\underline{t}} m_0(t)dt \leq \pi \int_{\underline{t}}^{1} m_1(t)dt + (1-\pi) \int_{0}^{\underline{t}} m_0(t)dt \\ &\iff \int_{0}^{\underline{t}} m_0(t)dt \geq 0. \end{aligned}$$

This shows there is no profitable deviation for the principal, to disclosing the state, when the state is 0.

Now we show that such a deviation does not exist even when the state is 1, under optimal mechanism of type F_-. If $\underline{t}$ is such that $m_1(\underline{t}) \geq 0$, by Claim D.5.1,

$m_1(t) \geq 0$ for all $t > \underline{t}$. Therefore $\int_{\underline{t}}^{1} m_1(t)dt \geq 0$ and we are done. Therefore for the rest of the proof we assume $m_1(\underline{t}) < 0$.

Let $\underline{\pi}$ and $\overline{\pi}$ be as defined in Theorem 2.A. We know when $V_0 \leq 0$, the optimal test is of type F_- for $\pi \leq \underline{\pi}$. Now consider the same problem, but for $\pi \geq \overline{\pi}$. By theorem 2.A, in this case the optimal test is of type F_+, with a corresponding signal threshold of, say $\bar{t}$. By identical reasoning as above, $\int_{\bar{t}}^{1} m_1(t)dt \geq 0$. By Claim D.4.5, $\bar{t} < \underline{t}$. By Claim D.5.1, for all $t' > \bar{t}$ such that $m_1(t') \leq 0$, $\int_{t'}^{1} m_1(t)dt \geq \int_{\bar{t}}^{1} m_1(t)dt \geq 0$. Hence $\int_{\underline{t}}^{1} m_1(t)dt \geq 0$.