Test Design under Falsification*

Eduardo Perez-Richet† Vasiliki Skreta‡

November 26, 2021

Abstract

We study the optimal design of tests with manipulable inputs. Tests take a unidimensional state of the world as input, and output an informative signal to guide a receiver's approve or reject decision. The receiver wishes to only approve states that comply with her baseline standard. An agent with a preference for approval can covertly falsify the state of the world at a cost. We characterize receiver-optimal tests and show they rely on productive falsification by compliant states. They work by setting a more stringent operational standard, and granting noncompliant states a positive approval probability to deter them from falsifying to the standard. We also study how falsification-detection technologies improve optimal tests. They allow the designer to build an implicit cost of falsification into the test, in the form of signal devaluations. Exploiting this channel requires enriching the signal space.

Keywords: Information Design, Falsification, Tests, Manipulation, Cheating, Persuasion.
JEL classification: C72; D82.

*We thank anonymous referees for excellent comments. Ricardo Alonso, Philippe Jehiel, Ines Moreno de Barreda, Meg Meyer, Philip Strack, Peter Sorensen, and Thomas Wiseman provided helpful comments and suggestions. Eduardo Perez-Richet acknowledges funding by the Agence Nationale de la Recherche (STRATCOM-16-TERC-0010-01) at early stages of this research, and by the European Research Council (ERC) consolidator grant 101001694 at later stages. Vasiliki Skreta acknowledges funding by the European Research Council (ERC) consolidator grant 682417 "Frontiers In Design." Angelos Diamantopoulos, Alkis Georgiadis-Harris, Richard Faltings, Amir Habibi, Nathan Hancart, and Ignacio Núñez provided excellent research assistance at various stages.

†Sciences Po, CEPR – e-mail: eduardo.perez@sciencespo.fr
‡UT Austin, UCL, CEPR – e-mail: vskreta@gmail.com

1 Introduction

In modern economies, decisions are increasingly guided by tests, ratings, and algorithms. Yet, these information-production technologies are vulnerable to input manipulations, that is, *falsification*. Consider, for instance, the problem of regulating vehicles' emissions. Compliance with emission standards must be checked by testing. However, emissions tests have proved to be manipulable through *defeat devices*[1] that artificially reduce vehicles' emissions in testing conditions. Accounting for possible input manipulations is an integral part of designing tests that provide valuable information. We propose a theory in which an agent can manipulate a test by *covertly* falsifying its inputs. We show optimal tests must induce *productive falsification*, that is, falsification that serves the interests of the designer. We also examine how the availability of a falsification-detection technology may improve optimal tests and affect their nature. We show enriching the set of signals and shaping the test so that signals are progressively devalued in proportion to the amount of falsification is then optimal, thereby allowing the designer to build an implicit cost of falsification into the test.

Our analysis is based on a model of test manipulation as costly falsification of inputs. We now motivate this choice with additional examples. Financial institutions may hide assets or misreport their holdings when facing stress tests. Teachers may teach their students to the test, effectively falsifying their true ability. Online shoppers may adapt their browsing behavior to get better deals from pricing algorithms.[2] Falsification costs may reflect expected fines or reputational damage in case manipulations are discovered, explicit financial or technological costs, psychological lying costs,[3] or the opportunity cost of altering one's behavior as in the online-shopping example. We examine the impact of their magnitude. We show that, whereas higher falsification costs benefit the designer, they have a non-monotonic effect on the agent's payoff.

[1] "... *with defeat devices programmed into the vehicles' complex emissions control software. The devices cause the vehicles to produce compliant results during emissions testing. But when not running a test, the vehicles' emissions controls perform differently, and less effectively...*" Source: The United States department of Justice; see also https://www.justice.gov/opa/press-release/file/1316601/download.

[2] As another example, the German artist Simon Weckert hacked the Google Maps algorithm for a performance, creating a fake traffic jam by walking a cart filled with cell phones along a street of Berlin.

[3] Evidence that lying is costly is documented in Abeler, Nosenzo, and Raymond (2019), for example.

We study a designer-agent-receiver model. A state of the world is drawn from a bounded interval that contains both positive and negative states. The designer, seeking to maximize the receiver's welfare, commits to a test (a Blackwell experiment) that takes the state of the world as an input, and outputs an informative signal. Based on this signal, the receiver makes a binary approve-reject decision. Her gain from approval is equated with the state of the world, so her *baseline standard* for approval is 0, and her first-best is to approve positive, henceforth *compliant*, states, and reject negative, henceforth *noncompliant*, states. The agent has a state-independent preference for approval. Knowing its design, he can covertly falsify the state of the world that goes into the test. We say that falsification is *productive* whenever it raises the approval probability of compliant states while preserving that of noncompliant ones compared to no falsification.

In the emissions example, the test designer is the regulator (the EPA in the US).[4] The state of the world is the difference between the emission standard and the true emission level. The EPA also acts as the receiver, deciding whether a vehicle conforms to environmental standards.[5] The agent is a car manufacturer, who can resort to defeat devices to falsify emission levels while being tested.

We assume the cost of falsification, $\gamma c(t|s)$, depends on the (true) *source* state s and the *target* state t, and is increasing in the distance between t and s. The scaling factor γ captures the magnitude of falsification costs. We make two additional assumptions: First, for noncompliant states, falsifying as the highest state is more costly than falsifying as the lowest state. Second, the cost function satisfies the triangular inequality for upward falsification. Given our monotonicity assumption, we can interpret the triangular inequality as a form of increasing returns to the scale of falsification.

Theorem 1 characterizes a receiver-optimal test. A *recommendation principle* allows us to focus on obedient tests with two signals, *approve* and *reject*. The optimal test recommends approval with *top approval probability* p for states above an *operational standard* $\hat{s}$, and with nominal probability $\{p - \gamma c(\hat{s}|s)\}^+$ for other

[4] "*All new cars and trucks sold in the U.S. must be certified to meet federal emission standards, such as limits on the amount of smog-forming and greenhouse gas emissions that they can produce. Most testing is performed by auto manufacturers at their own facilities. EPA then audits the data and performs its own testing on some of the vehicles to confirm the manufacturers' results.*" https://www.epa.gov/greenvehicles/testing-national-vehicle-and-fuel-emissions-laboratory

[5] We show the designer's problem and optimal outcome are identical if the designer can commit both to a test *and* a contingent approval rule (Proposition 1), so our results are valid whether or not the designer and the agent are the same entity, as long as they have aligned preferences.

states.[6] For every state below $\hat{s}$ with a positive nominal approval probability, the agent is then indifferent between two optimal falsification strategies: not falsifying, or falsifying to the standard $\hat{s}$. Then, breaking this indifference in the receiver's favor is optimal, requiring that noncompliant states do not falsify while compliant states below $\hat{s}$ falsify to $\hat{s}$. Falsification is then *productive* because it allows all compliant states to be approved with top probability p, whereas noncompliant states are approved with their nominal probability.

The optimal values of p and $\hat{s}$ depend on the magnitude of falsification costs. If falsification costs are high, the optimal outcome is obtained by setting $p = 1$ and the standard $\hat{s}$ so that $\gamma c(\hat{s}|0) = 1$, which is just high enough to deter all noncompliant states from falsifying to $\hat{s}$. Productive falsification by compliant states then implies they are all approved with certainty, whereas noncompliant states are rejected with certainty due to the high falsification cost, yielding the receiver's first-best. With intermediate falsification costs, setting the highest possible operational standard and $p = 1$ is optimal. All compliant states are then approved with certainty, but some noncompliant states must be approved with positive probability to deter them from falsifying to the standard. When falsification costs are low, setting the highest operational standard and approving some noncompliant states with positive probability is still optimal. But the top approval probability must also be reduced ($p < 1$) to avoid approving extremely low states with positive probability, leading compliant states to be rejected with positive probability.

The intuition underlying the optimal test is that, by assigning the top approval probability p only to compliant states above $\hat{s}$ and letting lower compliant states falsify to the standard, the designer minimizes the approval probability of noncompliant states. If, instead, the test directly assigned probability p to all compliant states, some noncompliant states would falsify and get approved with probability p. We show inducing productive falsification is in fact necessary for optimal testing, so the truth-telling implication of the revelation principle fails in our framework.

We proceed to examine the effect of falsification-detection technologies on test design. Sophisticated tests and algorithms may include falsification-detection capabilities. We can think of such tests as relying on additional inputs that indicate whether the agent is falsifying the state, and to what extent. We model these technologies by simply assuming they make the agent's falsification strategy observ-

[6] Throughout the paper, we denote $z^+ = \max\{z, 0\}$.

able to the receiver. Thus, we study optimal test design under *overt* falsification. Overtness endows the designer with a new lever in the form of signal devaluations. Indeed, because deviations from an anticipated falsification strategy are observed, they lead the receiver to adjust her expectation following each signal. Devaluations occur when the posterior mean following a signal is adjusted downward, possibly leading the receiver to switch from approval to rejection. By building the threat of devaluation into the test, the designer creates an implicit cost of falsification that makes deviations less attractive, and improves test performance. To take advantage of this devaluation channel, however, the designer must use more than two signals.

To illustrate this idea, we characterize an optimal test when the state-space is binary in Theorem 2, and show it uses a continuum of signals that get progressively devalued as the amount of falsification increases. This characterization is possible because, in the binary-state setting, a falsification-proofness principle akin to the truth-telling implication of the revelation principle holds.

We then go back to the continuous-state model, where neither the falsification-proofness principle nor the recommendation principle hold. In Proposition 9, we show how to improve on the test from Theorem 1 by adding a third signal that leverages the devaluation channel. We thus obtain a new test that relies on both productive falsification and devaluations. The gains allowed by falsification detection are most important when falsification costs are low. If falsification is costless, relying on such technologies is the only way to deliver useful information to the receiver.

Our analysis contributes to practical test design by conceptualizing two levers to improve test performance: productive falsification and devaluations. Our test is equivalent to a mechanism in the tradition of Myerson (1982), and the literature on mediation and communication equilibria (Aumann, 1974; Forges, 1986). Indeed, as the principal in Myerson (1982), our designer commits to a mapping that takes the agent's report as input and outputs messages to the receiver. In this literature, the revelation principle is twofold, combining a *truth-telling* (or *falsification-proofness*) *principle* and a *recommendation principle*. This contrasts with our framework where costly falsification causes the falsification-proofness principle to fail with more than two states, and overt falsification causes the recommendation principle to fail.[7] Hence, we contribute to mechanism design theory

[7]Without costly falsification or overtness, a Myersonian principal cannot achieve anything in

by deriving optimal mechanisms in situations where the revelation principle fails. We also contribute to the literature on mechanism design with costly reporting, or falsification, by providing the first (to our knowledge) characterization of an optimal mechanism that induces falsification. Lacker and Weinberg (1989) incorporate costly state falsification in a model of risk-sharing contracts and characterize optimal falsification-proof contracts, but also show they may be outperformed by contracts that induce falsification.[8]

Related literature. By introducing agency, in the form of costly state falsification, to the standard information design[9] setting of Kamenica and Gentzkow (2011) or Bergemann and Morris (2016), we add to a growing literature on information design when an agent can react to the experiment by undertaking an action that alters its informational content. For example, the agent can choose whether to take the test in Rosar (2017), or to disclose additional information in Bizzotto, Rüdiger, and Vigier (2020) and Terstiege and Wasser (2020).[10]

Frankel and Kartik (2021) and Ball (2020) study the optimal design of *linear scores* in a setting in which the agent has a privately known gaming ability (akin to our publicly known cost-scaling parameter γ) and the receiver has a continuum of actions and seeks to most accurately match the agent's fundamental type, which is the analog of our state of the world, and is multidimensional in Ball (2020). The logic of their results is that information about gaming ability tends to crowd out information about fundamental type. Under their assumptions, falsification does not distort information when gaming ability is public because higher types falsify higher. This is in stark contrast to our model, despite the fact that we study similar agency frictions. Another distinction is that we characterize optimal tests without restrictions on the class of tests the designer can choose from.

The two aforementioned papers build on Frankel and Kartik (2019), who study the effect of gaming without taking a design perspective. In an analogous vein, Hu, Immorlica, and Vaughan (2019) analyze strategic manipulations of a *given* classification algorithm, and Cunningham and Moreno de Barreda (2015) equilibrium

our framework (see Remark 1).

[8]The relatively small economics and computer science literature on mechanism design with reporting costs (Kephart and Conitzer, 2016; Deneckere and Severinov, 2017; Severinov and Tam, 2019) focuses on mechanisms with transfers. All these papers provide conditions on reporting costs to ensure truth-telling is without loss.

[9]See Bergemann and Morris (2019) and Kamenica (2019) for reviews of this literature

[10]Other examples include Lipnowski, Ravid, and Shishkin (2019) and Nguyen and Tan (2020), where the *agent* is the sender, who can manipulate the output of the experiment.

state falsification in a model with a fixed testing technology.

Falsification can be interpreted as lying, which connects our paper to the literature on strategic communication and interactions with costly lying (Kartik, Ottaviani, and Squintani, 2007; Kartik, 2009; Sobel, 2020). The key difference from these works is that we design optimal channels (tests) rather than relying on direct unmediated communication. Falsification can also be thought of as signalling (Spence, 1973), in a model in which each type of the agent (state of the world) corresponds to a distinct *natural* (least costly) action, and the test takes these actions as inputs. The agent might then find choosing a different action so as to influence the decision of the receiver is optimal. The cost of falsification is simply the opportunity cost of deviating from the natural action.

2 The covert-falsification model

A decision maker, henceforth *receiver*, can choose between two actions, which we label *approve* and *reject*. The receiver's payoff depends on a *state of the world*. She faces an *agent* with a state-independent preference for approval. The receiver can rely on information provided by a *test* that takes the state of the world as an input and outputs a signal. The agent can, however, manipulate the test by covertly *falsifying* the state of the world. We seek to solve the problem of a *designer* who can commit to a test so as to maximize the receiver's payoff.

States and payoffs. We normalize the receiver's rejection payoff to 0, and equate the state of the world $s \in S$ with her payoff from approval, where $S = \left[-\underline{s}, \overline{s}\right]$, and $-\underline{s} < 0 < \overline{s}$. We let $S^- = [-\underline{s}, 0)$ and $S^+ = [0, \overline{s}]$, and henceforth refer to states in S^- as *negative*, or *noncompliant*, and to states in S^+ as *positive*, or *compliant*. Thus, the receiver wishes to approve compliant states, and reject noncompliant states. We say 0 is the *baseline standard* for approval. The agent obtains payoff 1 upon approval, and 0 otherwise.

Prior. The prior distribution of states of the world is a probability measure π, which we assume to be atomless and have full support on S. We denote its cdf as F_π, and its mean as $\mu_\pi = \mathbb{E}_\pi(s)$. If $\mu_\pi < 0$, we let s_0 denote the largest state such that the receiver would approve if she knew all lower states are excluded. Hence, s_0 is the unique state such that $\mathbb{E}_\pi(s|s \geq s_0) = 0$. For convenience, we adopt the

convention that $s_0 = -\underline{s}$ when $\mu_\pi \geq 0$.

Tests. A test is a Blackwell experiment (Blackwell, 1951, 1953): a measurable space of signals X, and a Markov kernel τ from S to X, so that $\tau(s) \in \Delta X$ denotes the distribution of signals generated by state s. The prior π and the test τ together define a joint probability measure on $X \times S$ that we denote by $\tau\pi$.

Falsification. A falsification strategy ϕ is a Markov kernel from S to S. If T is a Borel subset of S and $s \in S$ a state of the world, $\phi(T|s)$ denotes the probability that the true state s, or *source*, is falsified as a *target state* in T. We denote by $\phi(s) \in \Delta S$ the distribution of falsified states generated by the true state s. The *truth-telling strategy* is the Markov kernel δ that maps each state s to the Dirac measure δ_s, which puts probability 1 on target state s. Together, the prior π and the falsification strategy ϕ define the joint probability measure $\phi\pi$ on $S \times S$.

Falsifying s as t comes at cost $\gamma c(t|s)$, where $c : S \times S \to \mathbb{R}_+$ is a measurable function, and $\gamma \geq 0$ is a scaling parameter that captures the magnitude of falsification costs. The cost of falsification strategy ϕ is then $C(\phi) = \gamma \int_{S\times S} c \, d\phi\pi$.

Information structures. Together, a falsification strategy ϕ and a test τ define an *information structure* embodied by the Markov kernel $\tau\phi : S \to X$, which, combined with the prior π, defines a joint distribution $\tau\phi\pi$ on $X \times S$. Then, $\tau\phi(s) \in \Delta X$ denotes the distribution of signals generated by state s. Note that, although $\tau\phi$ cannot be more Blackwell informative than τ, it is not necessarily less Blackwell informative. In particular, the receiver may prefer $\tau\phi$ to τ. This possibility plays an important role in our results as we find that optimal tests induce productive falsification by the agent.

Approval. The action space of the receiver is $A = \{a, r\}$, where a stands for *approval* and r for rejection. An approval strategy of the receiver is a Markov kernel α from X to A. We denote by δ^A the Markov kernel from A to itself that, to each action $a \in A$, assigns the Dirac measure δ_a, which puts probability 1 on a. If the signal space is $X = A$, we refer to δ^A as the *obedient* strategy for the receiver.

Outcome and expected payoffs. An *outcome* ω is a Markov kernel from S to A, which defines the approval probability of any state. Then, $\omega\pi$ is a joint

$$\begin{tikzcd} S \arrow[r, "\phi", "\text{Agent}"'] \arrow[rrr, bend right, "\omega"'] & S \arrow[r, "\tau", "\text{Designer}"'] & X \arrow[r, "\alpha", "\text{Receiver}"'] & A \end{tikzcd}$$

Figure 1: *Outcomes, tests, and strategies*

distribution on $A \times S$. Falsification costs aside, both players only care about outcomes. Specifically, the receiver's payoff is $V(\omega) = \mathbb{E}_{\omega\pi}(s|a)$, and the approval probability under ω is $\Pi(\omega) = \omega\pi(\{a\} \times S)$. Together, a test τ, a falsification strategy ϕ, and an approval strategy α determine an outcome $\alpha\tau\phi$. The agent's payoff is then $U(\alpha\tau\phi, \phi) = \Pi(\alpha\tau\phi) - C(\phi)$.

Timing. The timing of the game is as follows:

1. **Test:** A test τ is exogenously given and publicly observable.
2. **Falsification:** The agent covertly chooses a *falsification strategy* ϕ.
3. **State:** The state s is realized according to π.
4. **Testing and results:** The falsification strategy generates a falsified state of the world t according to $\phi(s)$, and the test generates a publicly observable signal x about the falsified state of the world according to $\tau(t)$.
5. **Receiver decision:** The receiver forms beliefs and chooses to approve or reject.

For convenience, we assume the agent chooses his falsification strategy *ex ante*, before the state is realized. However, this choice of timing is inconsequential because ex ante and *interim* falsification (knowing the state) are essentially equivalent.[11]

Solution concept and equilibrium. Our solution concept is perfect Bayesian equilibrium. A pair (ϕ, α) is an equilibrium under τ if (i) the receiver's posterior is derived using Bayes' rule given $\tau\phi$ whenever possible, (ii) the receiver approves optimally given her belief, and (iii) the agent's falsification strategy ϕ is optimal given the receiver's approval strategy.

[11] See Lemma S1.2 in the Online Appendix.

Posterior beliefs. For each signal x occurring with positive probability under $\tau\phi\pi$, a receiver anticipating ϕ forms a posterior belief in ΔS according to Bayes' rule whenever possible, that is, for every $x \in \bigcup_{s\in S} \operatorname{supp} \tau\phi(s)$, and arbitrarily otherwise. In both cases, we denote this belief by $\tau\phi\pi_x$. Let $\mu(x|\tau\phi) = \int_S s\, d\tau\phi\pi_x$ denote the associated posterior mean.

Receiver-optimality. Given τ, the approval strategy α of a receiver anticipating ϕ is optimal if and only if it satisfies $\alpha(\boldsymbol{a}|x) = 1$ if $\mu(x|\tau\phi) > 0$, and $\alpha(\boldsymbol{a}|x) = 0$ if $\mu(x|\tau\phi) < 0$. The receiver's value function only depends on the information structure, and we denote it by $\overline{V}(\tau\phi) = \max_\alpha V\big(\alpha\tau\phi\big)$.

Equilibrium feasibility. We say that a pair (τ, ϕ) is *equilibrium feasible*, or that ϕ is equilibrium feasible under τ, if an approval strategy α exists such that (ϕ, α) is an equilibrium under τ, that is, if and only if

$$\exists\alpha, \quad \begin{cases} \forall x,\ \mu(x|\tau\phi) > 0 \Rightarrow \alpha(\boldsymbol{a}|x) = 1, \\ \forall x,\ \mu(x|\tau\phi) < 0 \Rightarrow \alpha(\boldsymbol{a}|x) = 0, \\ \forall \phi',\ U(\alpha\tau\phi, \phi) \geq U(\alpha\tau\phi', \phi'). \end{cases} \tag{EF}$$

The designer's problem. We consider a test designer who seeks to maximize the receiver's payoff. His problem is then to find an information structure (τ, ϕ) that maximizes $\overline{V}(\tau\phi)$ subject to (EF). In the remainder of this paper, we refer to such an information structure as optimal. By extension, we also refer to the test τ as optimal.

Falsification costs. We assume the cost function satisfies some basic properties. First, truth-telling is costless, $c(s|s) = 0$. Second, it is monotonic in the sense that falsifying to and from states that are further away is strictly more costly. Formally, $c(t|s) < c(t'|s)$ for all s, t, t' such that $t' < t \leq s$ or $s \leq t < t'$; and $c(t|s) < c(t|s')$ for all s, s', t such that $s' < s \leq t$ or $t \leq s < s'$. Finally, it is continuous. We also make two more substantial assumptions that play an important role for our results.

Definition 1. *The cost function*

(i) has the costlier-to-top property if

$$c(\overline{s}|0) \geq \min\{c(-\underline{s}|0), 1\}; \tag{CTT}$$

(ii) satisfies the upward triangular inequality if, for every $s \leq m \leq t$,

$$c(t|m) + c(m|s) \geq c(t|s). \tag{UTI}$$

The *costlier-to-top* property says that falsifying from the baseline standard to the highest state is more costly than falsifying to the lowest state that is worth falsifying to. By monotonicity, this comparison extends to all noncompliant states. (CTT) thus captures in a relatively unrestrictive manner the intuitive idea that falsifying upward is more costly than falsifying downward.

The *upward triangular inequality* can be interpreted as putting a bound on the cost increase of falsifying further up, as $c(t|s) - c(m|s) \leq c(t|m)$. If the cost function is differentiable, it implies that the cost increase of falsifying to a marginally higher target state is bounded by the cost of a marginal falsification from the initial target: $c_t(t|s) \leq c_t(t|t)$. In particular, the cost of a marginal upward falsification must then be everywhere positive.

Consider a cost function such that, for $t \geq s$, $c(t|s) = f(s)g(t-s)$, where f is a positive-valued and g is a nonnegative-valued increasing function with $g(0) = 0$. Then, it satisfies (UTI) whenever g is concave, or more generally subadditive, and f is nondecreasing. Subadditivity then captures increasing returns to scale in the size of falsification.[12]

3 Test design under covert falsification

We start with two key preliminary results that simplify the analysis. First, we establish a *recommendation principle* that allows us to restrict attention to tests that equate signals to action recommendations. Second, we show ex-ante and interim falsification are essentially equivalent. We then solve the designer's problem.

[12]Note that convexity in the size of falsification can also be accommodated provided that the cost scaler increases sufficiently fast with the source as is the case with the cost function $c(t|s) = e^{2\beta s/\alpha}\{\alpha(t-s) + \beta(t-s)^2\}$ for $t \geq s$, where $\alpha > 0$ and $\beta \geq 0$.

3.1 Preliminary results

Recommendation principle. Mimicking standard results in Myerson (1982) and Kamenica and Gentzkow (2011), we establish a recommendation principle. According to this principle, if a test τ gives rise to an equilibrium (ϕ, α), it can equivalently be replaced by the garbled test $\alpha\tau$, with signal space $X = A$, that gives rise to an equilibrium consisting of the same falsification strategy ϕ for the agent and the obedient approval strategy δ^A for the receiver. Both equilibria are outcome equivalent since $\alpha\tau\phi = \delta^A(\alpha\tau)\phi$ and therefore lead to the same payoffs for both players. Whereas the result that obedience is a best response to ϕ under the new test is standard, the result that ϕ remains a best response to δ^A is specific to our setting, and leverages the fact that, in equilibrium, covert deviations from ϕ do not affect the receiver's decisions.[13]

For the remainder of our analysis, we therefore, in a slight abuse of notation, redefine tests as measurable functions $\tau : S \to [0, 1]$, where $\tau(s)$ is the probability that the test recommends approval in state s. We refer to this probability as the *nominal approval probability* of state s. Because $X = A$, the composition of a test τ and a falsification strategy ϕ defines an outcome $\omega = \tau\phi$. We say that $\omega = \tau\phi$ is an *equilibrium outcome* if (τ, ϕ) is equilibrium feasible. The *true approval probability*, henceforth also denoted by $\omega(s)$, may differ from the nominal probability.

With this redefinition, we can write the agent's payoff as

$$U(\tau\phi, \phi) = \int_{S\times S} \tau(t) d\phi\pi(t, s) - C(\phi),$$

and the receiver's payoff as

$$V(\tau\phi) = \int_{S\times S} s\tau(t) d\phi\pi(t, s).$$

Obedience requires the receiver's posterior mean following the *approve* signal to be nonnegative, $\int_{S\times S} s\tau(t) d\phi\pi(t, s) \geq 0$, and her posterior mean following the *reject* signal to be nonpositive, $\int_{S\times S} s(1 - \tau(t)) d\phi\pi(t, s) = \mu_\pi - \int_{S\times S} s\tau(t) d\phi\pi(t, s) \leq 0$. Piecing these two inequalities together, the obedience constraint puts a lower bound on the receiver's payoff, requiring that she obtains at least her payoff in the

[13] The formal statement (Lemma S1.1) and proof of this result are in the Online Appendix.

absence of information:

$$V(\tau\phi) \geq \max\{\mu_\pi, 0\}. \tag{RO}$$

Equivalence of ex-ante and interim falsification. Working with the recommendation principle, the receiver's obedience constraint takes care of the receiver's side of the equilibrium feasibility condition (EF), which can therefore be reduced to requiring optimality of the agent's falsification strategy ϕ:

$$\int_{S\times S} \big\{\tau(t) - \gamma c(t|s)\big\} d\phi\pi(t,s) \geq \int_{S\times S} \big\{\tau(t) - \gamma c(t|s)\big\} d\phi'\pi(t,s), \quad \forall \phi'. \tag{EF'}$$

If, instead, the agent chooses ϕ at the interim stage, after observing the state, the condition for ϕ to be interim equilibrium feasible is

$$\phi\big(\text{argmax}_t \ \tau(t) - \gamma c(t|s) \,|\, s\big) = 1, \quad \forall s. \tag{IEF}$$

Standard arguments show[14] (EF') is equivalent to the interim condition holding for almost every s. Because falsification from a subset of states with measure 0 has no effect on the players' ex-ante payoffs, we restrict attention to falsification strategies that satisfy (IEF).

Costless falsification. We briefly consider costless falsification ($\gamma = 0$) as a benchmark. In this case, the truth-telling implication of the revelation principle applies and, combined with (IEF), implies the test must give a constant approval probability to all states. By the recommendation principle, this probability must be 0 if $\mu_\pi < 0$, and 1 if $\mu_\pi > 0$.

Remark 1 (Costless falsification)**.** When falsification is costless, the unique equilibrium outcome is uninformative, and the receiver rejects if $\mu_\pi < 0$, and approves if $\mu_\pi > 0$. Her payoff is equal to $\max\{\mu_\pi, 0\}$. ◇

The designer's program. By the recommendation principle and interim-ante equivalence, we can find an optimal test by solving the following *designer's program*:

$$\sup_{\tau,\phi} \ V(\tau\phi) \quad \text{s.t.} \ \ (\text{IEF}), \ (\text{RO}) \tag{$\mathcal{P}$}$$

[14]For a formal statement and a proof, see Lemma S1.2 in the Online Apendix.

Next, we argue (RO) is redundant and can be relaxed without loss of generality. Indeed, a test with a constant nominal approval probability is uninformative and satisfies (IEF) because it makes falsification irrelevant for the agent. Furthermore, the uninformative test achieves the lower bound required by (RO). Any solution to the relaxed program

$$\sup_{\tau,\phi} \; V(\tau\phi) \quad \text{s.t.} \quad \text{(IEF)} \tag{$\mathcal{P}'$}$$

must give the receiver a higher payoff than the uninformative test, and therefore also satisfy (RO). Hence, it is also a solution to ($\mathcal{P}$).

Interestingly, this redundancy implies the designer does not benefit from more commitment power. Indeed, the program of a designer with the power to commit to an approval strategy of the receiver based on reports about the state, or to a test and an approval strategy together, is exactly ($\mathcal{P}'$).

Proposition 1 (Value of commitment)**.** *Commitment to an approval strategy, or to a test and an approval strategy, has no additional value than commitment to a test for the designer.*

The relaxed program ($\mathcal{P}'$) can also be interpreted as that of a principal seeking to allocate a good to an agent of type s, where s is the principal's payoff of allocating the good to the agent. The principal's payoff from the outside option (not allocating the good) is 0; the agent gets a state-independent payoff from getting the good. The principal commits to an allocation probability τ that depends on the agent's report and misreporting is costly. Indeed, in such a problem the principal maximizes $\int_{S\times S} s\tau(t)d\phi\pi(t,s)$ which is equal to $V(\tau\phi)$ subject to (IEF). This interpretation connects our analysis to the literature on the design of optimal allocation rules without transfers. Ben-Porath, Dekel, and Lipman (2014) solve such a problem by exploiting costly verification, whereas Kattwinkel (2019) exploits private information of the principal correlated with the agent's type. We exploit costly reporting costs.

3.2 An optimal test

We start by introducing a simple class of tests. We then show we can restrict attention to this class, and characterize the optimal test within this class. Finally, we study its properties.

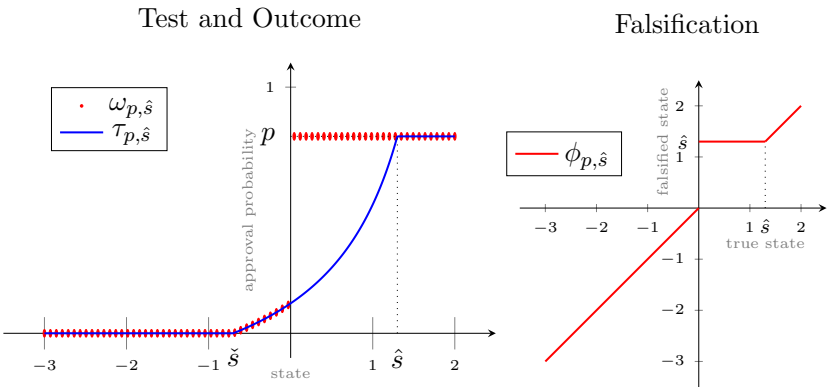

Figure 2: *On the left panel, the blue curve shows the nominal approval probabilities of $\tau_{p,\hat{s}}$, whereas the red dotted curve shows approval probabilities under the equilibrium outcome $\omega_{p,\hat{s}}$. On the right panel, the red curve illustrates the falsification strategy $\phi_{p,\hat{s}}$. The cost function is $\gamma c(t|s) = \frac{1.2|t-s|}{1+|t-s|}$ if $t \geq s$.*

An optimal class of tests. We consider a class of tests characterized by two parameters: a *top nominal approval probability* $p \in [0,1]$ and an *operational standard* $\hat{s} \in S^+$. A test sets the nominal approval probability of states above the operational standard to p and gives states below $\hat{s}$ the lowest nominal approval probability that deters them from falsifying to the standard:

$$\tau_{p,\hat{s}}(s) = \begin{cases} p & \text{if } s \geq \hat{s} \\ p - \gamma c(\hat{s}|s) & \text{if } s \in \left[\check{s}(p,\hat{s}), \hat{s}\right] \\ 0 & \text{if } s < \check{s}(p,\hat{s}) \end{cases},$$

where $\check{s}(p,\hat{s})$ is equal to the state $s \in S^-$ that solves $\gamma c(\hat{s}|s) = p$ when it exists. Otherwise, we set $\check{s}(p,\hat{s})$ equal to $-\underline{s}$.

Under (UTI), a test $\tau_{p,\hat{s}}$ makes truth-telling optimal in all states. To see why, we only need to consider the payoff of falsifying s as $t > s$, with $s, t \in \left[\check{s}(p,\hat{s}), \hat{s}\right]$[15]:

$$\tau_{p,\hat{s}}(t) - \gamma c(t|s) = p - \gamma c(\hat{s}|t) - \gamma c(t|s) \leq p - \gamma c(\hat{s}|s) = \tau_{p,\hat{s}}(s),$$

where the inequality is implied by (UTI).

By construction, $\tau_{p,\hat{s}}$ also makes the agent indifferent between falsifying to the operational standard and truth-telling for states in $\left[\check{s}(p,\hat{s}), \hat{s}\right]$. Therefore, the agent has multiple optimal falsification strategies. Among these, the designer

[15] Other cases follow from cost monotonicity, and the flatness of the test outside of this interval.

can break indifferences in favor of the receiver, requiring the agent to only falsify compliant states below $\hat{s}$ to the standard. Let

$$\phi_{p,\hat{s}}(s) = \begin{cases} \delta_{\hat{s}} & \text{if } s \in [0, \hat{s}] \\ \delta_s & \text{otherwise} \end{cases}$$

denote this strategy. The resulting outcome $\omega_{p,\hat{s}} = \tau_{p,\hat{s}}\phi_{p,\hat{s}}$ is that all compliant states are approved with top probability p, whereas noncompliant states are approved with their nominal approval probability, as illustrated in Figure 2. Formally,

$$\omega_{p,\hat{s}} = \begin{cases} p & \text{if } s \geq 0 \\ p - \gamma c(\hat{s}|s) & \text{if } s \in \big[\check{s}(p, \hat{s}), 0\big) \\ 0 & \text{if } s < \check{s}(p, \hat{s}) \end{cases}.$$

In summary, we have shown these are equilibrium outcomes, as stated in the following lemma.

Lemma 1. *If the cost function satisfies* (UTI)*, the falsification strategy* $\phi_{p,\hat{s}}$ *satisfies* (IEF) *under* $\tau_{p,\hat{s}}$*.*

Optimal test. Optimizing the receiver's payoff within the class of equilibrium outcomes $\{\omega_{p,\hat{s}}\}$ reduces the original infinite dimensional problem to a two dimensional one. Our next result, Theorem 1, characterizes an outcome $\omega_{p,\hat{s}}$ within our class that solves the designer's program ($\mathcal{P}$). To simplify the exposition, we only state the theorem in the case where $\mu_\pi < 0$, and refer the reader to Theorem 3 in the Appendix for a complete statement and a proof.

Theorem 1. *Suppose the cost function satisfies* (UTI) *and* (CTT)*. Then,* $\big(\tau^*_\gamma, \phi^*_\gamma\big)$ *solves* ($\mathcal{P}$)*, where* $\tau^*_\gamma = \tau_{p^*_\gamma \hat{s}^*_\gamma}$*,* $\phi^*_\gamma = \phi_{p^*_\gamma \hat{s}^*_\gamma}$*,* $\hat{s}^*_\gamma = \max\big\{s \in S : \gamma c(s|0) \leq 1\big\}$*, and* $p^*_\gamma = \min\big\{\gamma c(\overline{s}|s_0), 1\big\}$*.*

We denote the optimal equilibrium outcome by $\omega^*_\gamma = \tau^*_\gamma \phi^*_\gamma$. The shape of the optimal test and outcome are illustrated in Figure 3. We can divide the range of γ into three regions as follows.

In the *high-cost region*, $\gamma \geq 1/c(\overline{s}|0)$, setting $p^*_\gamma = 1$, and $\hat{s}^*_\gamma > 0$ to solve

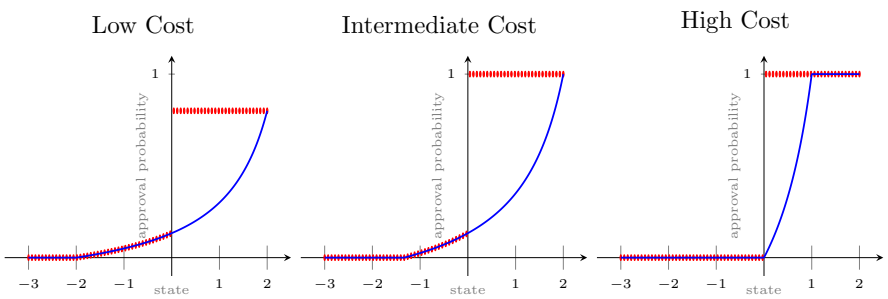

Figure 3: *Optimal test and outcome in the different regions, where* $\pi = U([-3, 2])$ *so* $s_0 = -2$, $c(t|s) = \frac{|t-s|}{1+|t-s|}$ *if* $t \geq s$, *and* $\gamma \in \{1, 1.3, 2\}$.

$\gamma c(\hat{s}^*|0) = 1$ is optimal, so that

$$\tau^*_\gamma(s) = \begin{cases} 1 & \text{if } s \geq \hat{s}^*_\gamma \\ 1 - \gamma c(\hat{s}^*_\gamma|s) & \text{if } s \in \left[0, \hat{s}^*_\gamma\right] \\ 0 & \text{if } s < 0 \end{cases}.$$

The optimal outcome is the receiver's first-best $\omega^*_\gamma(s) = \mathbb{1}_{s \geq 0}$, so noncompliant states are rejected and compliant states approved with certainty. To reach first-best, the designer only needs to raise the operational standard $\hat{s}^*_\gamma$ above the baseline standard, and let the agent do the correction by falsifying. Indeed, a test that recommends rejection below and approval above $\hat{s}^*_\gamma$, both with certainty, also yields the optimal outcome[16] ω^*_γ.

In the *intermediate-cost region*, $1/c(\overline{s}|s_0) \leq \gamma < 1/c(\overline{s}|0)$, setting $p^*_\gamma = 1$ and $\hat{s}^*_\gamma = \overline{s}$ is optimal, so that

$$\tau^*_\gamma(s) = \left\{1 - \gamma c(\overline{s}|s)\right\}^+,$$

with corresponding equilibrium outcome:

$$\omega^*_\gamma(s) = \begin{cases} 1 & \text{if } s \geq 0 \\ \gamma\left\{1 - c(\overline{s}|s)\right\} & \text{if } s \in \left[\check{s}(1, \overline{s}), 0\right) \\ 0 & \text{if } s < \check{s}(1, \overline{s}) \end{cases}.$$

Hence, optimality requires setting the highest possible operational standard. Compliant states are approved with certainty, but some noncompliant states must be

[16]See our discussion on multiplicity of optimal tests and their characterization below.

approved with positive probability to deter them from falsifying to the standard.

In the *low-cost region*, $\gamma < 1/c(\overline{s}|s_0)$, setting $p^*_\gamma = \gamma c(\overline{s}|s_0)$ and $\hat{s}^*_\gamma = \overline{s}$ is optimal, so that

$$\tau^*_\gamma(s) = \begin{cases} \gamma\big\{c(\overline{s}|s_0) - c(\overline{s}|s)\big\} & \text{if } s \in \big[s_0, \overline{s}\big] \\ 0 & \text{if } s < s_0 \end{cases},$$

with corresponding equilibrium outcome:

$$\omega^*_\gamma(s) = \begin{cases} \gamma c(\overline{s}|s_0) & \text{if } s \geq 0 \\ \gamma\big\{c(\overline{s}|s_0) - c(\overline{s}|s)\big\} & \text{if } s \in \big[s_0, 0\big) \\ 0 & \text{if } s < s_0 \end{cases}.$$

As in the intermediate-cost region, optimality requires setting the highest possible operational standard and approving some noncompliant states with positive probability. But now, it also requires rejecting compliant states with positive probability to deter very low states (below s_0) from falsifying to the standard. To illustrate this trade-off, consider using instead the test $\tau_{p,\overline{s}}$, with $p = p^*_\gamma + \varepsilon$, for a small $\varepsilon > 0$. Under this test, the true approval probability of all states above s_0 increases by ε, leading to a null gain as $\varepsilon \int_{s_0}^{\overline{s}} s dF_\pi(s) = 0$. But the receiver also incurs a strict loss over states below s_0, as some of those states are approved with positive probability.

Characterization of optimal tests. Our optimal test is not unique. The optimal equilibrium outcome, however, is essentially unique. Furthermore, we can characterize the set of optimal tests that do not penalize the agent relative to the test of Theorem 1. To see why our optimal test is not unique, consider two types of variations. First, we can lower the nominal approval probability of compliant states below $\hat{s}^*_\gamma$ without changing the agent's equilibrium falsification strategy or the outcome. Indeed, this operation only strengthens the incentive of these states to falsify to the standard.[17] Second, when the standard is not set to the highest state (in the high-cost region), we can also lower the nominal approval probability of states above $\hat{s}^*_\gamma$ so as to make them falsify (downward) to the standard, without changing the equilibrium outcome or the receiver's payoff. However, this operation

[17] To complete the argument, we show the modification cannot incentivize the agent to change his falsification strategy in any other way.

lowers the agent's payoff, because he needs to falsify more. If we rule out optimal tests that unnecessarily penalize the agent, only variations of the first type are possible. Variations of this type are in some sense more robust since they can make the incentive for productive falsification strict. See Proposition S1.1 in the Online Appendix for a formal statement and a proof.

A corollary of this characterization is that productive falsification is necessary for optimality. Optimal tests that do not penalize the agent must induce essentially the same falsification strategy. Other optimal tests induce even more falsification.

Theorem 1: proof overview. We next provide a sketch of the proof, which can be found in its entirety in the Appendix. Working with the relaxed program ($\mathcal{P}'$), the main step to prove Theorem 1 consists in showing that for every equilibrium outcome $\omega = \tau\phi$, we can find an outcome $\omega_{p,\hat{s}}$ that makes the receiver better off.

Consider such an equilibrium outcome $\omega = \tau\phi$. Assume, for simplicity, ϕ is degenerate and the function τ admits a unique maximizer $\hat{s}_\tau = \text{argmax}_s\, \tau(s)$ on S, and let $p = \tau(\hat{s}_\tau)$. Then, we can distinguish three cases, depending on the location of $\hat{s}_\tau$. To provide intuition about the role of the (CTT) assumption, we next discuss two of these cases.

Suppose first that $\hat{s}_\tau \geq 0$ and $\gamma c(\hat{s}_\tau|0) \leq p$. Then, we set the standard $\hat{s} = \hat{s}_\tau$. Under the outcome $\omega_{p,\hat{s}}$, each compliant state is approved with probability p, which is at least as high as under ω. However, some noncompliant states are approved with positive probability. Let $s \in [\check{s}(p, \hat{s}), 0)$ be such a state. Let $t(s) = \text{argmax}_{t'}\, \tau(t') - \gamma c(t'|s)$ be an optimal falsification target for s under τ, so $\omega(s) = \tau(t(s))$. Then, by optimality of $t(s)$,

$$\omega(s) \geq \tau(t(s)) - \gamma c(t(s)|s) \geq \tau(\hat{s}) - \gamma c(\hat{s}|s) = p - \gamma c(\hat{s}|s) = \omega_{p,\hat{s}}(s).$$

So, s is approved with lower probability under $\omega_{p,\hat{s}}$ than under ω.

Suppose next that $\hat{s}_\tau < 0$ and $\gamma c(\hat{s}_\tau|0) \leq p$. Then, we choose the standard $\hat{s} > 0$ such that $c(\hat{s}|0) = c(\hat{s}_\tau|0)$. (CTT) ensures that doing so is possible. Under $\omega_{p,\hat{s}}$, each compliant state is approved with probability p, which is higher than under ω. As in the former case, consider a noncompliant state s approved with positive probability under ω. Then,

$$\omega(s) = \tau\big(t(s)\big) \geq \tau(t(s)) - \gamma c(t(s)|s) \geq \tau(\hat{s}_\tau) - \gamma c(\hat{s}_\tau|s) \geq p - \gamma c(\hat{s}|s) = \omega_{p,\hat{s}}(s),$$

where the second inequality is due to the optimality of falsifying as $t(s)$, and the third inequality is due to cost monotonicity. Again, the approval probability of noncompliant states is lowered under $\omega_{p,\hat{s}}$.

3.3 Properties of optimal tests

We discuss the shape of the optimal test, its welfare properties, and comparative statics with respect to the cost parameter γ. We state our results under the assumption $\mu_\pi < 0$, but it is easy to adapt the results.[18]

Comparative statics and asymptotics. The receiver's payoff under the optimal outcome is

$$V_\gamma^* = \int_S s\omega_\gamma^*(s)dF_\pi(s).$$

Because the agent is indifferent between ϕ_γ^* and truth-telling, we can evaluate his payoff as if he were using the truth-telling strategy; hence,

$$U_\gamma^* = \int_S \tau_\gamma^*(s)dF_\pi(s).$$

Proposition 2 (Comparative statics)**.** *V_γ^* is increasing in γ in the low, and intermediate-cost regions, but constant and equal to the full-information payoff in the high-cost region. U_γ^* is increasing in γ in the low, and high-cost regions, and decreasing in the intermediate-cost region.*

It is natural that the receiver's payoff increases as falsification becomes more costly. The agent's payoff, however, is non-monotonic in the cost. To see why, note the cutoff state $\check{s}_\gamma^* = \check{s}(p_\gamma^*, \hat{s}_\gamma^*)$ at which the nominal approval probability starts increasing is fixed to 0 in the high-cost region, and to s_0 in the low-cost region. Therefore, a steeper cost function (higher γ) leads to higher nominal approval probabilities for all states above this cutoff. In the intermediate-cost region, by contrast, the top approval probability is fixed to 1, and it is awarded exclusively to the highest state, whereas the positive probability cutoff $\check{s}_\gamma^*$ increases with γ. A steeper cost function therefore leads to decreasing the nominal approval probabilities of all states. The next result considers limit tests and payoffs, and

[18] When $\mu_\pi \geq 0$, the low cost region does not exist, but the comparative statics of Proposition 2 is otherwise unchanged. The only difference in Proposition 3 is that the uninformative payoffs are 1 for the agent and μ_π for the receiver.

its proof is immediate by taking limits in γ for the optimal test and outcome functions.

Proposition 3 (Asymptotics)**.** *Both the outcome and the test converge to the uninformative test as $\gamma \to 0$. As $\gamma \to \infty$, the outcome converges to $\mathbb{1}_{s\geq 0}$ and the test to $\mathbb{1}_{s>0}$. Payoffs converge accordingly, to the uninformative payoffs in the first case: $\lim_{\gamma\to 0} U^*_\gamma = \lim_{\gamma\to 0} V^*_\gamma = 0$, and to the full-information payoffs in the latter: $\lim_{\gamma\to\infty} V^*_\gamma = \mathbb{E}_\pi(s|s \geq 0)$ and $\lim_{\gamma\to\infty} U^*_\gamma = \mathbb{P}_\pi(s \geq 0)$.*

Welfare. Falsification is a friction that generates inefficiencies. Our optimal outcome is *constrained efficient* by definition, because it maximizes the receiver's payoff under falsification. However, it is never unconstrained efficient, and the welfare loss generated by falsification can be decomposed into two channels: First, a *direct loss* due to the cost of productive falsification by the agent; second, an *informational loss* arising indirectly from distortions the designer needs to build into the outcome to optimally manage the falsification friction.

We measure total welfare loss, the direct falsification loss, and the informational loss as follows. First, we equate the direct loss to the total falsification cost incurred by the agent $C(\phi^*_\gamma)$. The agent's payoff net of this cost is his expected approval probability $\Pi^*_\gamma = \mathbb{E}_\pi\big(\omega^*_\gamma(s)\big)$, so by restoring the falsification cost to the agent, we reach the point $\big(V^*_\gamma, \Pi^*_\gamma\big)$ in the payoff space. Starting from this point, we measure the informational loss as the sum of payoff gains to both players that can be obtained by moving to the closest point on the unconstrained Pareto frontier. To do so, we start by measuring the payoff gain the receiver could obtain by freely reorganizing approval probabilities according to an outcome function ω', while keeping the expected approval probability of the agent constant $\Pi(\omega') = \Pi^*_\gamma$. Because the receiver prefers to concentrate the probability of approval on higher states, a solution to this reorganization problem is the threshold function $\omega'(s) = \mathbb{1}_{s\geq\tilde{s}}$ for $\tilde{s} \geq s_0$ such that $\mathbb{P}_\pi(s \geq \tilde{s}) = \Pi^*_\gamma$. This reorganization might lead to an approval threshold $\tilde{s} > 0$ if the approval probability Π^*_γ is too low, which is the case for low values of γ. Then, choosing instead $\tilde{s} = 0$ leads to higher payoff gains for both players, and we measure the informational loss as the sum of these gains. To summarize, we measure total welfare loss as:

$$WL = \underbrace{C(\phi^*_\gamma)}_{\text{direct loss}} + \underbrace{V(\omega') - V^*_\gamma + \Pi(\omega') - \Pi^*_\gamma}_{\text{informational loss}},$$

where $\omega' = \mathbb{1}_{s \geq \tilde{s}}$ and $\tilde{s} = \max\{s \in [s_0, 0] : \mathbb{P}_\pi(s \geq \tilde{s}) \geq \Pi^*_\gamma\}$.

This decomposition implies that our constrained optimal outcome suffers from an informational loss in the low, and intermediate-cost regions but not in the high-cost region. The direct loss, however, is always present. It is increasing in the low, and intermediate-cost regions, decreases in the high-cost region, and asymptotically vanishes as falsification becomes arbitrarily costly.

Figure 4 illustrates both the comparative statics and asymptotic behavior of payoffs, as well as the welfare loss due to falsification. The grey area depicts the set of feasible payoffs in the absence of falsification. As it shows, the falsification cost borne by the agent can be heavy: in the high-cost region (III), the agent may lose more than half his full-information payoff in falsification cost, while the receiver still benefits from her full-information payoff.

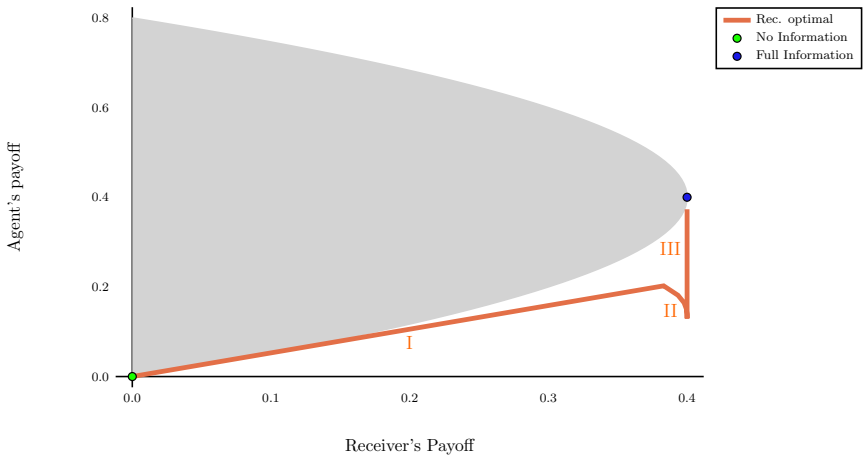

Figure 4: *The grey area depicts the set of attainable payoffs under all possible information structures in the absence of falsification. The orange path shows the payoffs from the optimal test as a function of γ. The curve starts at the no-information payoffs for $\gamma = 0$, moves successively across the low-cost region (I), the intermediate-cost region (II) and the high-cost region (III), and heads toward the full-information payoffs as γ increases. $\gamma : 0 \to 5$; $c(t|s) = |t - s|/(1 + |t - s|)$, if $t \geq s$; $\pi = Uniform(-3, 2)$.*

4 Test design with falsification detection

We seek to understand how the availability of a falsification-detection technology affects test design. To focus on the effect of detection in its purest form, we assume a technology that perfectly reveals the falsification strategy of the agent to the receiver, so that falsification is *overt* rather than *covert*. The timing of the

game is the same, but the receiver now learns the agent's falsification strategy ϕ before choosing her action. Her posterior beliefs therefore reflect actual rather than anticipated falsification.[19] In most of this section, to simplify the exposition, we assume *upward-only falsification*: the agent can only falsify to higher states.[20]

A few remarks are in order. First, the optimal equilibrium of Theorem 1 remains an equilibrium in the overt case.[21] Thus, the ability to detect falsification does not hurt the receiver, and indeed, the same test remains optimal in the high-cost region where it attains first-best. We show it can be improved when the cost is lower. Second, neither the recommendation principle nor the ante-interim equivalence hold any longer, making the analysis of the overt case substantially more difficult. Third, the result of Proposition 1 on the value of more commitment no longer holds: if the designer can commit to the receiver's action, committing to reject whenever falsification is present delivers the first-best outcome.

Intuitively, falsification detection provides the designer with a new lever in the form of signal devaluations. Indeed, deviations from equilibrium by the agent lead the receiver to revise the posterior mean associated with a given signal downward (*devaluation*), or upward (*appreciation*), and adjust her action accordingly. By ensuring deviations induce detrimental devaluations, the designer can therefore impose implicit devaluation costs to the agent in addition to the explicit falsification costs. In this section, we show how the designer can use these implicit devaluation costs to improve on the best equilibrium outcome of the covert case.

We proceed as follows. To address the technical difficulties, we first study a binary-state version of our model. This simplifies the analysis by the availability of a falsification-proofness principle that allows us to restrict attention to tests that the agent has no incentive to falsify. In this setup, we characterize an optimal test relying on the idea of devaluations. Our characterization shows using the devaluation channel requires a rich signal space: although adding a third signal is sufficient to allow the designer to get a significant improvement from the devaluation effect, optimality requires using tests with a granular signal space, even in our simple binary-state binary-action framework. We then go back to our initial model with a continuum of states, and show how to use ideas from the binary-state

[19] For a full definition of equilibrium under overt falsification, see Online Appendix S2.

[20] A condition on downward falsification costs ensuring our results hold when downward falsification is possible always exists. We state this condition explicitly for the binary-state case in Proposition 8.

[21] It also remains optimal in the covert case under upward-only falsification, as we show in Theorem S1.1 of the Online Appendix.

model to improve on the optimal test of the covert case from Theorem 1 when falsification costs are low.

4.1 Falsification detection in the binary-state model

The binary-state model. In this model, the state space is $S = \{-\underline{s}, \overline{s}\}$. Slightly abusing notation, we denote by π the prior probability of the high state $\pi(\overline{s})$. We assume $\mu_\pi = \pi\overline{s} - (1-\pi)\underline{s} < 0$, and we let $\varphi_0 = \frac{\pi\overline{s}}{(1-\pi)\underline{s}}$ denote the probability with which the low state needs to be pooled with the high state to bring the expectation attached to the pool to 0.[22] We let $\underline{\phi} = \phi(\overline{s}|-\underline{s})$, $\overline{\phi} = \phi(-\underline{s}|\overline{s})$, $\underline{c} = \gamma c(\overline{s}|-\underline{s})$, and $\overline{c} = \gamma c(-\underline{s}|\overline{s})$. With upward-only falsification, a falsification strategy is simply defined by the probability $\underline{\phi}$, and truth-telling corresponds to $\underline{\phi} = 0$.

Fully-informative and binary-signal tests. To gain intuition, consider first a fully informative test with $X = \{\underline{x}, \overline{x}\}$ and $\tau(\overline{x}|\overline{s}) = \tau(\underline{x}|-\underline{s}) = 1$. Suppose $\underline{c} < 1$. Following the high signal $\overline{x}$, the receiver's expected payoff from approval is $\frac{\pi\overline{s}-(1-\pi)\underline{\phi}\underline{s}}{\pi+(1-\pi)\underline{\phi}}$, so she approves if $\underline{\phi} \leq \varphi_0$. Because the agent can only falsify upward, the receiver is certain the state is $-\underline{s}$ after $\underline{x}$, and rejects. The agent's payoff is therefore equal to $\left\{\pi+\underline{\phi}(1-\pi)(1-\underline{c})\right\} \mathbb{1}_{\underline{\phi}\leq\varphi_0}$, so he optimally chooses $\underline{\phi} = \varphi_0$, which is the falsification level that makes the receiver indifferent between both actions when receiving the high signal. The resulting information structure is the one the agent would design if given the opportunity (as in Kamenica and Gentzkow, 2011). It is agent-optimal and receiver-pessimal. The receiver's payoff is zero, as without any information. When falsification is costless, the agent obtains his first-best payoff. As the falsification cost increases, the agent's payoff falls, but the test and the receiver's payoff remain unchanged. Note this is in fact the best outcome for both the agent and the receiver under *any* binary-signal test.

A three-signal test. Under overt falsification, enriching the test with additional signals can make the receiver better off. The intuition is that additional signals allow the designer to get more traction from the devaluation effect. We next illustrate this idea with a three-signal test that dominates all binary-signal tests. As a corollary, it proves the recommendation principle no longer holds in the overt case.

[22] φ_0 is analogous to s_0 in the continuous-state model.

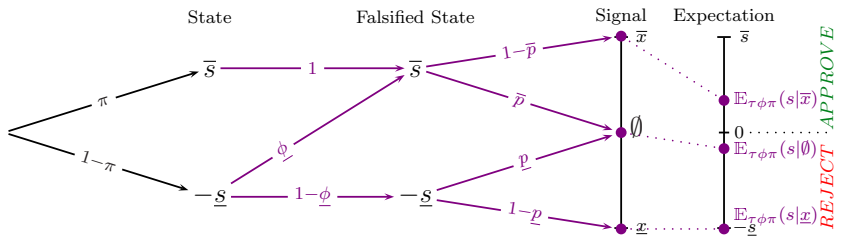

Figure 5: *Three-signal test: The expectation column shows the devaluation effect of upward falsification on posterior means.*

Consider a test with discrete signal space $X = \{\underline{x}, \emptyset, \overline{x}\}$, and such that $\tau(\overline{s})$ is the probability distribution $(0, \overline{p}, 1-\overline{p})$, and $\tau(-\underline{s}) = (1-\underline{p}, \underline{p}, 0)$, as illustrated in Figure 5. We set $\underline{p}/\overline{p} = \varphi_0$, so that, in the absence of falsification:

$$\mathbb{E}_{\tau\pi}(s|\overline{x}) = \overline{s}, \qquad \mathbb{E}_{\tau\pi}(s|\emptyset) = 0, \qquad \mathbb{E}_{\tau\pi}(s|\underline{x}) = -\underline{s},$$

leading the receiver to approve after $\emptyset$ and $\overline{x}$, and reject otherwise. With upward-only falsification, for any $\underline{\phi} > 0$, we have:

$$\mathbb{E}_{\tau\phi\pi}(s|\overline{x}) \propto \big(\pi\overline{s} - \underline{\phi}(1-\pi)\underline{s}\big), \;\; \mathbb{E}_{\tau\phi\pi}(s|\emptyset) \propto \underline{\phi}\big(\pi\overline{s} - (1-\pi)\underline{s}\big) < 0, \;\; \mathbb{E}_{\tau\phi\pi}(s|\underline{x}) = -\underline{s}.$$

Therefore, any amount of falsification triggers the devaluation of signal $\emptyset$, leading the receiver to reject. The agent trades off this implicit cost of falsification against the benefit of increasing the probability that signal $-\underline{s}$ generates signal $\overline{x}$. If the agent chooses $\underline{\phi} > 0$, he must ensure $\mathbb{E}_{\tau\phi}(s|\overline{x}) \geq 0$ so the receiver approves after $\overline{x}$, implying $\underline{\phi} \leq \varphi_0$. The agent's payoff for $0 \leq \underline{\phi} \leq \varphi_0$ is therefore

$$\pi(1-\overline{p}) + \pi\overline{p}\,\mathbb{1}_{\underline{\phi}=0} + (1-\pi)\underline{\phi}\big\{1 - \overline{p} - \underline{c}\big\}.$$

Hence, setting $\overline{p} \geq \frac{\overline{s}(1-\underline{c})}{\underline{s}+2\overline{s}}$ ensures the agent has no incentive to falsify. The receiver is then certain the state is compliant when she gets the high signal and is strictly better off under this test than with no information or any binary-signal test. Furthermore, the receiver is better off with smaller values of $\underline{p}$ (and hence $\overline{p}$), because it lowers her probability of approving noncompliant states. Therefore, the best test she can pick in this class is obtained by setting $\overline{p} = \frac{\overline{s}(1-\underline{c})}{\underline{s}+2\overline{s}}$. With this test, the receiver obtains $\frac{\underline{s}+(1+\underline{c})\overline{s}}{\underline{s}+2\overline{s}}\pi\overline{s}$, which is strictly positive even if $\underline{c} = 0$.

Pushing the intuition that additional signals are key to leveraging devaluations, the optimal test we derive next uses a continuum of signals. The reader can now

either proceed to Section 4.2 where we characterize the optimal test in the binary-state model, or proceed directly to Section 4.3 where we derive a three signal test that simultaneously relies on devaluation and productive falsification in the continuum of states setting.

4.2 Optimal testing in the binary-state model

A falsification-proofness principle. In the binary-state case, we can rely on a revelation-principle type of result allowing us to restrict attention to tests that induce truth-telling as an equilibrium falsification strategy.[23] To understand why it holds in the binary-state case, suppose a falsification strategy ϕ is equilibrium feasible under test τ. Then, consider the alternative test $\tau' = \tau\phi$. Any information structure $\tau'\phi'$ attainable under τ' can be attained under τ by using the falsification strategy $\phi\phi'$, generating the same best-response from the receiver in each case. However, in the binary-state case, $C(\phi') \geq C(\phi\phi') - C(\phi)$, implying ϕ' can be a profitable deviation from truth-telling under τ' only if $\phi\phi'$ is a profitable deviation from ϕ under τ, a contradiction. Therefore, τ' yields an equilibrium under which the agent does not falsify and the receiver obtains the same payoff as under $\tau\phi$. Note that, in contrast to the usual revelation principle, the payoff of the agent is higher under $\tau'\delta$ than under $\tau\phi$ because he saves $C(\phi)$. The receiver's payoff and the outcome are identical.

Normalizing signals as means. As in much of the information design literature, we can use the mean-based (or, equivalently, in the binary-state case, belief-based) approach to simplify our problem.[24] We thus describe tests by the distribution of posterior means they generate, which amounts to normalizing signals as means. A test is therefore represented as a distribution of posterior means with cdf H over $[-\underline{s}, \overline{s}]$ with the *martingale property* that $\int_{-\underline{s}}^{\overline{s}} x dH(x) = \mu_\pi$, which is equivalent to (integrating by parts)

$$\int_{-\underline{s}}^{\overline{s}} H(x)dx = \overline{s} - \mu_\pi. \tag{MP}$$

[23]We establish this principle in Proposition S2.1 of the Online Appendix. It holds under either overt or covert falsification. With more than two states, the cost inequality $C(\phi') \geq C(\phi\phi') - C(\phi)$ may fail: if ϕ falsifies m to t and ϕ' falsifies s to m, with $s < m < t$, then $\phi\phi'$ must falsify both s and m to t.

[24]See Lemma S2.1 in the Online Appendix for a formal treatment.

As in Kolotilin (2018) and Gentzkow and Kamenica (2016), this test can be equivalently represented by the function $\mathcal{H}(x) = \int_{-\underline{s}}^{x} H(y)dy$ from $[-\underline{s}, \overline{s}]$ to $[0, \overline{s} - \mu_\pi]$, which is nondecreasing and convex, with $\mathcal{H}(-\underline{s}) = 0$ and $\mathcal{H}(\overline{s}) = \overline{s} - \mu_\pi$. Let Δ^B denote the set of nondecreasing convex functions from $[-\underline{s}, \overline{s}]$ to $[0, \overline{s} - \mu_\pi]$ that satisfy these properties. This representation is known to be without loss of generality in the absence of falsification. With falsification, we need to show that pooling together all signals leading to the same posterior mean does not modify the falsification incentives of the agent. Using this representation, we hereafter equate signals with the posterior mean they generate given the test (and in the absence of falsification).

Rewriting Payoffs. Under test $\mathcal{H}$, and in the absence of falsification ($\underline{\phi} = 0$), the receiver's payoff is[25]

$$V(\mathcal{H}, 0) = \int_0^{\overline{s}} x dH(x) = \mu_\pi + \mathcal{H}(0)$$

and the agent's payoff is

$$U(\mathcal{H}, 0) = 1 - H_\ell(0),$$

where $H_\ell(x) = \lim_{\substack{y \to x \\ y < x}} H(y)$ is also the left derivative of $\mathcal{H}$ at x and gives the probability of generating a posterior mean strictly below x.

Equilibrium characterization. Increasing $\underline{\phi}$ sends the noncompliant state toward any positive signal x at a higher rate, thus lowering the posterior mean formed by the receiver when observing x. If x is sufficiently close to 0, this devaluation leads the receiver to no longer approve x. In effect, falsification results in a new threshold signal $\hat{x}(\underline{\phi})$ such that the receiver only approves for signals $x \geq \hat{x}(\underline{\phi})$. Interestingly, this threshold depends on falsification only: it is independent of the test.

Lemma 2. *If $\underline{\phi} > \varphi_0$, all signals lead to rejection. If $\underline{\phi} \leq \varphi_0$, a threshold $\hat{x}(\underline{\phi}) = \frac{-\mu_\pi \underline{s}\underline{\phi}}{\pi(\overline{s}+\underline{s}) - \underline{\phi}\underline{s}}$ exists such that the receiver approves for signals $x \geq \hat{x}(\underline{\phi})$, and rejects otherwise.*

Lemma 2 implies falsification levels outside of $[0, \varphi_0]$ are dominated for the agent. Furthermore, because a one-to-one relationship exists between any falsifi-

[25] The second expression for the receiver's payoff is obtained using integration by parts.

cation level $\underline{\phi}$ in this range and the threshold it generates on $[0, \overline{s}]$, we can reformulate the agent's falsification problem as the choice of an approval threshold[26] $x \in [0, \overline{s}]$ for the receiver, induced by falsification level:

$$\hat{\phi}(x) = \frac{(\underline{s} + \mu_\pi)x}{(x - \mu_\pi)\underline{s}}.$$

Proposition 4 (Equilibrium characterization)**.** *Given a test $\mathcal{H}$, an equilibrium is characterized by an approval threshold $x \in [0, \overline{s}]$ for the receiver, and a falsification level $\underline{\phi} \in [0, \varphi_0]$ such that $\underline{\phi} = \hat{\phi}(x)$, and x maximizes the agent's payoff:*

$$U\big(\mathcal{H}, \hat{\phi}(x)\big) = 1 - \left(1 + \frac{x}{\underline{s}}\right) H_\ell(x) + \frac{x}{\underline{s}(x - \mu_\pi)}\mathcal{H}(x) - \frac{(1-\pi)(\underline{s} + \mu_\pi)x}{(x - \mu_\pi)\underline{s}}\underline{c}.$$

The only part of the proposition that needs an explanation is the calculation of the agent's payoff. Given the prior, falsification level, and threshold, we only need to know the distributions of signals generated by each of the two states $\overline{s}$ and $-\underline{s}$ to perform this computation. Their cdfs are, respectively,[27]

$$\overline{H}(x) = \frac{1}{\mu_\pi + \underline{s}}\big\{(x + \underline{s})H(x) - \mathcal{H}(x)\big\} \tag{$\overline{\text{CDF}}$}$$

and

$$\underline{H}(x) = \frac{1}{\overline{s} - \mu_\pi}\big\{(\overline{s} - x)H(x) + \mathcal{H}(x)\big\}. \tag{$\underline{\text{CDF}}$}$$

The designer's program. Using the falsification-proofness principle, we can formulate the designer's program as that of choosing a test function $\mathcal{H} \in \Delta^B$ to maximize $\mathcal{H}(0)$, under the falsification-proofness constraint that the agent has no incentive to induce any falsification threshold other than 0:

$$\begin{aligned} &\max_{\mathcal{H} \in \Delta^B} \mathcal{H}(0) \\ &\quad \text{s.t. } U\big(\mathcal{H}, 0\big) \geq U\big(\mathcal{H}, \hat{\phi}(x)\big), \qquad \forall x \in [0, \overline{s}]. \end{aligned} \tag{FPIC}$$

[26] In a slight abuse of notation, we denote this threshold by x, because every nonnegative signal can be induced as a threshold by some falsification strategy.

[27] To understand these expressions, note the joint probability that the state is compliant and the signal below x can be written both as $\pi\overline{H}(x)$ and as $\int_{-\underline{s}}^{x} \beta(z)dH(z)$, where $\beta(z) = \frac{z+\underline{s}}{\overline{s}+\underline{s}}$ is the updated probability of the compliant state conditional on having received signal z, and must therefore satisfy $\beta(z)\overline{s} - (1 - \beta(z))\underline{s} = z$. Integration by parts leads to the final formula.

Using the expression of the agent's payoff in Proposition 4, the constraint becomes:

$$H_\ell(x) - \frac{x}{(\underline{s}+x)(x-\mu_\pi)}\mathcal{H}(x) \geq \frac{\underline{s}}{\underline{s}+x}H_\ell(0) - \frac{\theta \underline{c} x}{(x-\mu_\pi)(\underline{s}+x)}, \qquad \forall x \in [0, \overline{s}] \tag{FPIC'}$$

where $\theta = (\overline{s} - \mu_\pi)(\underline{s} + \mu_\pi)/(\underline{s} + \overline{s})$.

Next, we derive a solution to the designer's program in two steps. First, we show we can restrict attention to tests that generate a single negative signal, or equivalently to tests such that $\mathcal{H}$ is linear over negative signals. Second, we show distributing positive signals so as to make the agent indifferent across all undominated falsification levels is optimal, or, equivalently, making the incentive constraint of the agent (FPIC') bind everywhere.

Linearization for negative signals. First, we can focus on test functions $\mathcal{H}$ that are linear on $[-\underline{s}, 0]$. Indeed, for any test function $\mathcal{H} \in \Delta^B$ that satisfies (FPIC'), the test function

$$\tilde{\mathcal{H}}(x) = \begin{cases} \frac{\mathcal{H}(0)}{\underline{s}}(x+\underline{s}) & \text{if } x \leq 0 \\ \mathcal{H}(x) & \text{if } x > 0 \end{cases}$$

is in Δ^B, delivers the same payoff to the receiver as $\mathcal{H}$, a higher payoff to the agent because $\tilde{H}_\ell(0) = \mathcal{H}(0)/\underline{s} \leq H_\ell(0)$ by convexity of $\mathcal{H}$, satisfies (FPIC') by the same argument, and is linear below 0.

Going back to the interpretation of test functions, this linearization implies we can focus on tests that generate a single negative signal equal to $-\underline{s}$. This signal is generated only by the low state.

Making the agent indifferent. Next, we characterize the unique test function that is linear below 0 and makes the agent indifferent across all thresholds induced by undominated falsification levels. By linearity, we can denote its slope below 0 by $\kappa \geq 0$, which is also the size of the atom it places on the negative signal. Our test function must then solve the indifference differential equation[28]

$$H(x) - \frac{x}{(\underline{s}+x)(x-\mu_\pi)}\mathcal{H}(x) = \frac{\kappa \underline{s}}{\underline{s}+x} - \frac{\theta \underline{c} x}{(x-\mu_\pi)(\underline{s}+x)} \tag{IDE}$$

[28]Note the subscript ℓ is no longer needed, because writing that H_ℓ satisfies this equality implies it is continuous, and therefore, $H_\ell = H$.

on $[0, \overline{s}]$, with initial condition $\mathcal{H}(0) = \kappa \underline{s}$. This linear differential equation has a unique solution parameterized by κ. For this solution to be a test function, it must satisfy the martingale property $\mathcal{H}(\overline{s}) = \overline{s} - \mu_\pi$, which pins down κ to a value that we denote by $\kappa^*_{\underline{c}}$, yielding the unique test function

$$\mathcal{H}^*_{\underline{c}}(x) = \kappa^*_{\underline{c}}(x + \underline{s}) + \left(\kappa^*_{\underline{c}}(\mu_\pi + \underline{s}) - \theta \underline{c}\right) \left\{ \left(\frac{x - \mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi + \underline{s}}} \left(\frac{x + \underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi + \underline{s}}} - 1 \right\} \mathbb{1}_{x>0},$$

where

$$\kappa^*_{\underline{c}} = \frac{\overline{s} - \mu_\pi + \theta \underline{c} \left\{ \left(\frac{\overline{s} - \mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi + \underline{s}}} \left(\frac{\overline{s} + \underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi + \underline{s}}} - 1 \right\}}{\overline{s} - \mu_\pi + (\underline{s} + \mu_\pi) \left(\frac{\overline{s} - \mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi + \underline{s}}} \left(\frac{\overline{s} + \underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi + \underline{s}}}}.$$

An optimal test. We show $\mathcal{H}^*_{\underline{c}}$ is in fact optimal.

Theorem 2. *$\mathcal{H}^*_{\underline{c}}$ is the unique test function that solves (IDE) on $[0, \overline{s}]$, and it solves the designer's problem under upward-only falsification.*

To understand why, note that in the class of partially linear tests we identified, the receiver's payoff depends on the size κ of the atom on the unique rejected signal $-\underline{s}$, which is only generated by the low state. $\mathcal{H}^*_{\underline{c}}$ puts an atom of size $\kappa^*_{\underline{c}}$ on this signal, and makes the agent indifferent across all approval thresholds he could induce through falsification. Increasing the size of this atom implies violating the falsification proofness condition for at least one falsification-induced threshold. For intuition, note that if $\mathcal{H}$ is a test that puts an atom of size $\kappa > \kappa^*_{\underline{c}}$ on the rejected signal, a signal x' between 0 and $\overline{s}$ must exist such that $\mathcal{H}$ first crosses $\mathcal{H}^*_{\underline{c}}$ from above at x' (if nowhere else, (MP) implies the two curves cross at $\overline{s}$). Furthermore, the left derivative $H_\ell(x')$ must be lower than $H^*_{\underline{c}}(x')$. However, combined with the fact that $\mathcal{H}^*_{\underline{c}}$ makes the agent indifferent across all thresholds, this inequality implies the agent prefers inducing falsification threshold x' to not falsifying under $\mathcal{H}$.

Properties of the test $\mathcal{H}^*_{\underline{c}}$. The following proposition derives some key properties of our optimal test. We depict its conditional and unconditional cdfs and densities in Figure 6.

Proposition 5 (Properties of CDF and PDF)**.** *$H^*_{\underline{c}}$ has support $\{-\underline{s}\} \cup [0, \overline{s}]$, with atoms at $-\underline{s}$ and $\overline{s}$, and a positive, continuously differentiable, and decreasing*

density on $[0, \overline{s})$. $\overline{H}^*_{\underline{c}}$ *has support* $[0, \overline{s}]$, *with a positive, continuously differentiable, and decreasing density on* $[0, \overline{s})$, *and a single atom at* $\overline{s}$. $\underline{H}^*_{\underline{c}}$ *has support* $\{-\underline{s}\} \cup [0, \overline{s}]$, *with a single atom at* $-\underline{s}$, *and a positive, continuously differentiable, and decreasing density on* $[0, \overline{s})$. *Furthermore,* $\overline{H}^*_{\underline{c}}$ *first-order stochastically dominates* $\underline{H}^*_{\underline{c}}$.

In spite of the binary-state and binary-action environment, the optimal test has a continuum of positive signals, and a single negative signal. A clustering of signals occurs close to 0 as illustrated in Figure 6. Furthermore, the test makes the agent indifferent across all undominated falsification levels[29] as it satisfies (IDE).

Granularity of positive signals, as well as the shape of the test, which is dictated by indifference, contribute to maximizing the implicit falsification cost at every falsification level. Increasing $\underline{\phi}$ devalues positive signals up to a threshold that does not depend on the test. When a signal is missing, the falsification level that would make this signal the new approval threshold is strictly dominated. By putting weight on such a signal, the designer can increase the associated implicit falsification cost, and at the same time lower the probability that the noncompliant state generates positive signals, thus increasing the receiver's payoff.

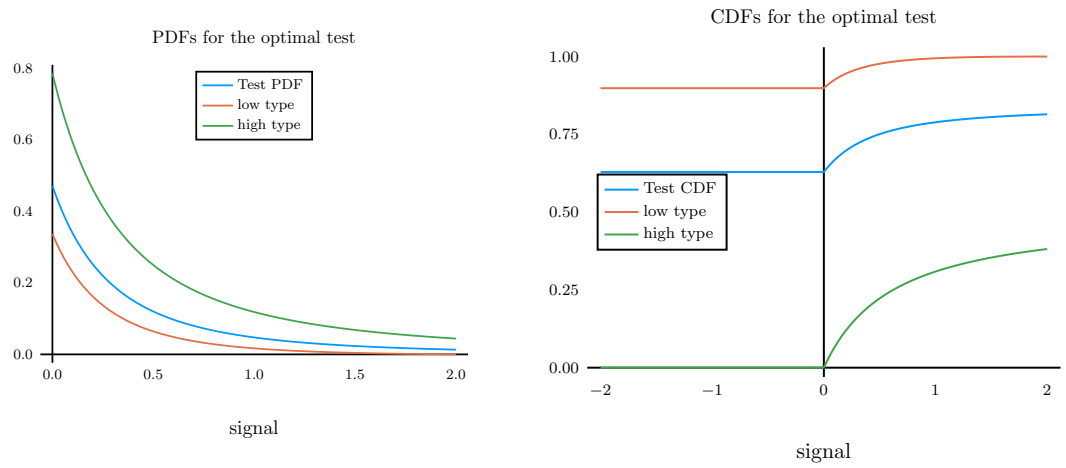

Figure 6: *PDF and CDF of the optimal test under overt falsification.* $-\underline{s} = -2$, $\overline{s} = 2$, $\pi = 0.3$.

Next, we examine the effect of falsification costs on payoffs. In contrast to the covert case, higher falsification costs are always detrimental to the agent and beneficial to the receiver.

[29] Indifference of the "agent" at the optimal information structure also appears in Roesler and Szentes (2017) and Ortner and Chassang (2018).

Proposition 6 (Comparative statics)**.** *The optimal test $\mathcal{H}^*_{\underline{c}}$ is increasing in $\underline{c}$ in the Blackwell informativeness order, and converges to the fully informative test function as $\underline{c} \to 1$. The receiver's payoff is also increasing in $\underline{c}$. The agent's payoff is decreasing in $\underline{c}$. Furthermore, $\mathcal{H}^*_{\underline{c}}$ is more Blackwell informative than any other optimal test function at $\underline{c}$.*

We proceed to discuss the welfare properties of the optimal test. In contrast to the covert case, the falsification friction does not induce any inefficiency. Indeed, falsification-proofness implies the absence of a direct loss, and there is no informational loss, because the compliant state is approved with certainty. Note that in the binary-state model, the direct loss is also absent in the covert case by the falsification-proofness principle, but informational inefficiencies persist when the falsification cost is low.[30]

Proposition 7 (Welfare)**.** *The optimal test $\mathcal{H}^*_{\underline{c}}$ is unconstrained efficient. It delivers at least half of the receiver's payoff under full information, and this bound is tight when $\underline{c} = 0$.*

The optimal test restores at least half of the receiver's full information payoff, even if falsification is costless. This is again in stark contrast with the covert case, as the receiver can then get no information at all under costless falsification. Next, we provide a necessary and sufficient condition on costs for $\mathcal{H}^*_{\underline{c}}$ to remain optimal when both upward and downward falsification are allowed.

Proposition 8 (Robustness to upward falsification)**.** *With both upward and downward falsification, constants $A > 0$ and B exist such that the test $\mathcal{H}^*_{\underline{c}}$ is optimal if and only if $A\overline{c} + B\underline{c} \geq 1$.*

To understand this result, note first that deviating to a falsification strategy $(\underline{\phi}, \overline{\phi})$ such that $\underline{\phi} + \overline{\phi} \leq 1$ is dominated by the strategy $(\underline{\phi}, 0)$, because it leads the receiver to use a threshold $\hat{x} \geq \hat{x}(\underline{\phi})$, while lowering the probability that the compliant state generates passing signals. Since $(\underline{\phi}, 0)$ is, by construction, unprofitable, $(\underline{\phi}, \overline{\phi})$ is also unprofitable. Therefore, we only need to show that under the condition of the proposition, deviations such that $\underline{\phi} + \overline{\phi} > 1$ are also non-profitable. The best of these deviations is such that $\underline{\phi} = 1 - \varphi_0$ and $\overline{\phi} = 1$. It gives the agent his best possible approval probability $\pi + (1 - \pi)\varphi_0$, at cost

[30] We provide a comparison of attainable payoffs under covert and overt falsification in the binary-state model in Online Appendix S3.

$\pi\overline{c} + (1-\pi)(1-\varphi_0)\underline{c}$. By comparing this payoff with the truth-telling payoff $1 - \kappa^*_{\underline{c}}$, we obtain the condition of the proposition.

4.3 Falsification detection in the continuous-state model

To illustrate how devaluations help the designer in the continuous-state model, we focus on the low-cost region where the optimal outcome in the covert case, ω^*_γ from Theorem 1, mandates both rejecting compliant states and approving some noncompliant states with positive probability. We construct a sequence of three-signal tests that rely on devaluations, and mirror the three-signal test from Section 4.1. However, they are modified to accommodate the continuum of states and, more importantly, to leverage productive falsification as τ^*_γ. We show outcomes from this sequence converge to an outcome under which compliant states are approved with certainty, and the receiver is better off than in the covert case. In contrast to the covert case, in which the optimal test is uninformative when falsification is costless, the tests we construct provide useful information to the receiver even under costless falsification.

We assume $\mu_\pi < 0$ and $\gamma c(\overline{s}|s_0) < 1$, so that we are in the low-cost region. We work with the signal space $X = \{\underline{x}, \emptyset, \overline{x}\}$. For each sufficiently small $\varepsilon > 0$, and each $p < 1 - \gamma c(\overline{s}|0)$, we define the test $\hat{\tau}_{p,\varepsilon}$ as follows (see Figure 7):

- $\hat{\tau}_{p,\varepsilon}(\emptyset|\overline{s}) = p$ and $\hat{\tau}_{p,\varepsilon}(\overline{x}|\overline{s}) = 1 - p$,
- For all $s \in [0, \overline{s})$, $\hat{\tau}_{p,\varepsilon}(\emptyset|s) = p$ and $\hat{\tau}_{p,\varepsilon}(\underline{x}|s) = 1 - p$,
- For all $s \in [-\underline{s}, s_0 - \varepsilon)$, $\hat{\tau}_{p,\varepsilon}(\underline{x}|s) = 1$,
- For all $s \in [s_0 - \varepsilon, 0)$, $\hat{\tau}_{p,\varepsilon}(\emptyset|s) = r_\varepsilon(p)$ and $\hat{\tau}_{p,\varepsilon}(\underline{x}|s) = 1 - r_\varepsilon(p)$, and $r_\varepsilon(p)$ satisfies
$$r_\varepsilon(p)\left(-\int_{s_0-\varepsilon}^{0} s dF_\pi(s)\right) = p\int_0^{\overline{s}} s dF_\pi(s). \tag{M0}$$

In the absence of falsification, $\overline{x}$ leads to a positive posterior mean, and $\underline{x}$ to a negative posterior mean, whereas (M0) ensures that $\emptyset$ leads to mean 0 and implies $r_\varepsilon(p) < p$. The receiver then rejects compliant states in $[0, \overline{s})$ with probability $1 - p$. But this inefficiency is overcome by productive falsification.

Productive falsification. The test $\hat{\tau}_{p,\varepsilon}$ (see part A of Figure 7) ensures the agent prefers falsifying all states $s \in [0, \overline{s})$ as $\overline{s}$, because it increases their approval

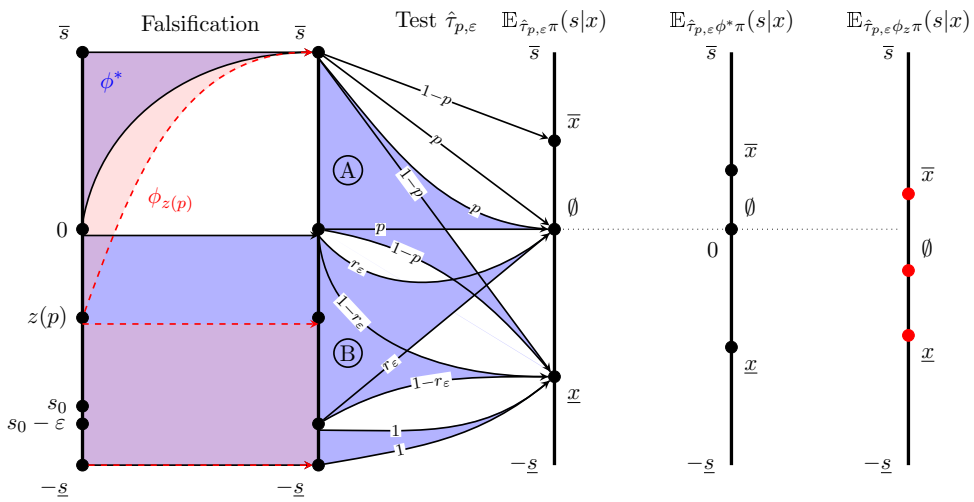

Figure 7: *Incentives for productive falsification are built into part A of the test, whereas the threat of devaluation is built into part B and ensures that ϕ^* is optimal. Under ϕ^*, all compliant states (productively) falsify as $\overline{s}$. Signal $\emptyset$ is devalued when noncompliant states are falsified as $\overline{s}$, as with $\phi_{z(p)}$. This is illustrated in the last column which shows posterior means under $\phi_{z(p)}$.*

probability by $1-p > \gamma c(\overline{s}|0) > \gamma c(\overline{s}|s)$. Let ϕ^* denote this falsification strategy. Under ϕ^*, $\overline{x}$ still leads to a positive posterior mean, whereas $\emptyset$ still leads to a posterior mean of 0. Hence, all compliant states are approved with certainty, whereas noncompliant states above $s_0-\varepsilon$ are approved with probability $r_\varepsilon(p)$, and lower noncompliant states are rejected with certainty. Next, we find conditions such that the threat of devaluation ensure ϕ^* is indeed the agent's best response.

Devaluation. For ϕ^* to be a best response, devaluation must in particular dissuade the agent from falsifying noncompliant states as $\overline{s}$. If a mass of noncompliant states above $s_0-\varepsilon$ falsify as $\overline{s}$, it increases the rate at which they generate $\emptyset$ as $p > r_\varepsilon(p)$, leading the receiver to form a negative posterior mean following $\emptyset$. Hence, falsifying noncompliant states as $\overline{s}$ bears both the explicit falsification cost, and the implicit falsification cost of devaluating signal $\emptyset$. Because any amount of falsification by noncompliant states leads to this devaluation, the best strategy of the agent that falsifies noncompliant states is then to falsify noncompliant states as much as the explicit cost allows as long as the posterior mean associated with $\overline{x}$ remains positive. The optimal deviation from ϕ^* is therefore to falsify as $\overline{s}$ all states (compliant and noncompliant) between $z(p) = \min\{s \geq s_0 : \gamma c(\overline{s}|s) \leq 1-p\}$ and $\overline{s}$ (strategy $\phi_{z(p)}$ in Figure 7). It has the benefit of increasing the probability

that noncompliant states generate $\overline{x}$, at the cost of devaluing signal $\emptyset$. Ensuring the agent is worse off under this strategy than under ϕ^*, and that ϕ^* is therefore a best response to $\hat{\tau}_{p,\varepsilon}$, puts a lower bound p_ε on p. Indeed, the gain over ϕ^* from this strategy is

$$\int_{s_0}^{0} \big\{1 - p - \gamma c(\overline{s}|s)\big\}^+ dF_\pi(s) - p\big[1 - F_\pi(0)\big] - r_\varepsilon(p)\big[F_\pi(0) - F_\pi(s_0 - \varepsilon)\big],$$

where the first term captures the new payoff for noncompliant falsifying as $\overline{s}$, whereas the two remaining terms capture the loss from devaluation. This gain is decreasing in p, so a minimal value p_ε of p exists ensuring a negative gain. The lower bound p_ε is then the value of p that makes the gain equate 0.

Because $r_\varepsilon(p)$ is increasing in p, lower values of p ensure a higher payoff for the receiver. To maximize the receiver's payoff while ensuring ϕ^* is an equilibrium response, we therefore choose p to be equal to the lower bound p_ε. This choice provides us with the family of tests $\hat{\tau}_{p_\varepsilon,\varepsilon}$ and equilibrium outcomes $\hat{\omega}_{p_\varepsilon,\varepsilon} = \hat{\tau}_{p_\varepsilon,\varepsilon}\phi^*$. The next proposition formally establishes these claims and shows the receiver's payoff from these outcomes increases as ε decreases to 0. Furthermore, the limit payoff dominates the receiver's payoff from ω^*_γ.

Proposition 9. *For every sufficiently small $\varepsilon > 0$, $\hat{\omega}_{p_\varepsilon,\varepsilon}$ is an equilibrium outcome. The receiver's payoff $V(\hat{\omega}_{p_\varepsilon,\varepsilon})$ is decreasing in ε. Furthermore, in the low-cost region, her limit payoff is higher than under ω^*_γ, that is $\lim_{\varepsilon\to 0} V(\hat{\omega}_{p_\varepsilon,\varepsilon}) > V^*_\gamma$.*

The test $\hat{\tau}_{p_\varepsilon,\varepsilon}$ improves performance by relying both on productive falsification and devaluations. Compared to τ^*_γ, we add a middle signal fated for approval. This operation provides incentives for productive falsification by making every state above $s_0 - \varepsilon$, except the top state, randomly generate either the low or the middle signal with almost equal probabilities when $\varepsilon \to 0$. Compliant states then have a strict incentive to falsify to the top state to be approved with certainty. The middle signal is constructed to make the receiver indifferent between his two actions when only compliant states falsify, but to be devalued whenever some noncompliant states falsify to the top. This devaluation effect is achieved by giving noncompliant states above $s_0 - \varepsilon$ a marginally lower probability of generating $\emptyset$ than compliant states. Note the limit test with $\varepsilon = 0$ mutes the devaluation effect because compliant and noncompliant states then generate $\emptyset$ with the same probability, so it is important that $\varepsilon > 0$. The limit outcome of these tests can be arbitrarily closely approximated, and it is that compliant states are approved

with certainty and noncompliant states above s_0 with uniform probability $p_0 = \lim_{\varepsilon \to 0} p_\varepsilon$, whereas lower states are rejected with certainty.

5 Conclusion

In the emissions cheating scandal, falsification by car manufacturers was detrimental as it enabled vehicles with noncompliant emission levels to pass the environmental test. Our analysis suggests that tests designed without accounting for falsification perform poorly when falsification is possible. In our model, the receiver-optimal test without falsification recommends approval for all compliant states, and rejection for noncompliant ones. Under falsification, however, this test induces detrimental falsification by noncompliant states sufficiently close to the baseline standard.

Our results point to practical and simple features that can significantly improve the performance of emissions (and other) tests. Under covert falsification, the structure of the optimal test in Theorem 1 suggests raising the operational standard above the baseline standard. A test with a high standard, on one hand deters detrimental falsification, whereas on the other hand is relies on productive falsification to generate approvals of compliant states. With high falsification costs simply raising the standard suffices to eliminate approvals of noncompliant states. With lower falsification costs, optimality additionally requires randomly approving a fringe of noncompliant states to deter detrimental falsification. When falsification costs are even lower, randomly rejecting compliant states becomes necessary to prevent extremely low states from falsifying to the standard.

When a falsification-detection technology is available, Theorem 2 and Proposition 9 show the threat of devaluation provides a powerful channel to improve test performance, which is especially appealing when falsification costs are low. In practice, a testing agency could accompany test outputs with a report on detected amounts of falsification, or even perform the devaluation on the receiver's behalf by directly reporting the expectation she should form following each output. A rich set of test outputs is key to harnessing this tool. Adding only a few signals might already yield strong benefits in practice. Indeed, in the binary-state case, a numerical analysis shows the three-signal test of Section 4.1 delivers at least 80% of the value of the optimal test to the receiver. Although we did not explicitly model the case of imperfect detection technologies, intuition suggests the

devaluation lever should remain operational in this context.

Appendix

Proof of Theorem 1. We prove a more general version of the theorem that also covers the case $\mu_\pi \geq 0$.

Theorem 3. *Suppose the cost function satisfies* (UTI) *and* (CTT)*. Then,* $(\tau^*_\gamma, \phi^*_\gamma) = (\tau_{p^*_\gamma, \hat{s}^*_\gamma}, \phi_{p^*_\gamma, \hat{s}^*_\gamma})$ *solves* $(\mathcal{P})$*, where*

(i) $\hat{s}^*_\gamma = \max\{s \in S : \gamma c(s|0) \leq 1\}$ *and* $p^*_\gamma = \min\{\gamma c(\overline{s}|s_0), 1\}$ *if* $\mu_\pi < 0$.

(ii) $\hat{s}^*_\gamma = \max\{s \in S : \gamma c(s|0) \leq 1\}$ *and* $p^*_\gamma = 1$ *if* $\mu_\pi \geq 0$.

Proof. We first show that for every pair (τ, ϕ) that satisfies (IEF), an outcome $\omega_{p,\hat{s}}$ exists that makes the receiver better off. Then, we optimize the receiver's payoff within this class. The proof follows the outline given in the paper, but accounts for the possibility of τ being discontinuous.

Step 1: Optimality of Class. Suppose $\omega = \tau\phi$ is an equilibrium outcome. Let $p = \sup_{s \in S} \tau(s)$, which exists because $\tau(\cdot)$ is bounded. For every $\varepsilon > 0$, let $S(\varepsilon) = \{s \in S : \tau(s) \geq p - \varepsilon\}$, and let $\bar{S}(\varepsilon)$ be the closure of $S(\varepsilon)$. By definition of p, each $S(\varepsilon)$, and hence, each $\bar{S}(\varepsilon)$, is nonempty. Furthermore, $\bar{S}(\varepsilon)$ is clearly nonincreasing in ε for the inclusion order. Therefore, by Cantor's intersection theorem, $\bar{S} = \bigcap_{\varepsilon > 0} \bar{S}(\varepsilon)$ is a nonempty compact subset of S.

If some $s \in S^+$ exists such that $\gamma c(s|0) \geq p$, we can set $\hat{s} \in S^+$ to be the unique state such that $\gamma c(\hat{s}|0) = p$. Then, under the outcome $\omega_{p,\hat{s}}$, every compliant state is approved with probability p, whereas every noncompliant state is rejected with certainty, making the receiver as least as well off as under ω. Otherwise, (CTT) implies $\gamma c(s|0) \leq p$ for every $s \in S$. Then, we consider two cases:

First, suppose $\bar{S} \cap S^+ \neq \emptyset$. This set is then a nonempty compact set, and we let $\hat{s}$ be its minimal element. Then, under $\omega_{p,\hat{s}}$, every compliant state is approved with probability p, which is at least as high as under $\tau\phi$. Next, we show noncompliant states pass with lower probability under $\omega_{p,\hat{s}}$. To see why, let $\{t_n\}$ be a sequence of nonnegative states that converges to $\hat{s}$ and such that the sequence $p_n = \tau(t_n)$ converges to p. Such a sequence exists because $\hat{s} \in \bar{S} \cap S^+$. Then, for every noncompliant state s, and every n, $\sup_t \ \tau(t) - \gamma c(t|s) \geq p_n - \gamma c(t_n|s)$, and going to the limit in n implies

$$\omega(s) \geq \sup_t \ \tau(t) - \gamma c(t|s) \geq p - \gamma c(\hat{s}|s) = \omega_{p,\hat{s}}(s).$$

Otherwise, we must have $\bar{S} \subset S^-$, and then we let $\tilde{s} = \max \bar{S} < 0$ and let $\hat{s} > 0$ be the unique compliant state such that $c(\hat{s}|0) = c(\tilde{s}|0)$, which must exist by (CTT). Then again, under $\omega_{p,\hat{s}}$, every compliant state is approved with probability p, which is at least as high as under ω. Next, we show noncompliant states pass with lower probability under $\omega_{p,\hat{s}}$. To see why, let $\{t_n\}$ be a sequence of states that converges to $\tilde{s}$ and such that the sequence $p_n = \tau(t_n)$ converges to p. Such a sequence exists because $\tilde{s} \in \bar{S}$. Then, for every noncompliant state s, and every n, $\sup_t \ \tau(t) - \gamma c(t|s) \geq p_n - \gamma c(t_n|s)$, and going to the limit in n implies

$$\omega(s) \geq \sup_t \ \tau(t) - \gamma c(t|s) \geq p - \gamma c(\tilde{s}|s) \geq p - \gamma c(\hat{s}|s) = \omega_{p,\hat{s}}(s).$$

Because, in each case, noncompliant states are approved with lower probability, and compliant states with higher probability, the receiver is better off under $\omega_{p,\hat{s}}$.

Step 2: Choosing parameters optimally. Let $V_{p,\hat{s}} = V(\omega_{p,\hat{s}})$ denote the receiver's payoff from an equilibrium outcome in our class. We distinguish four parameter regions and a change of p or $\hat{s}$ that increases the receiver's payoff in each of these regions. Together, these four operations imply the optimal values for $\hat{s}$ and p given in the theorem.

First, suppose $\mu_\pi < 0$ and $\gamma c(\hat{s}|s_0) < p$. Then, setting $p' = \gamma c(\hat{s}|s_0)$ is strictly better. Indeed,

$$V_{p',\hat{s}} - V_{p,\hat{s}} = (p' - p) \underbrace{\int_{s_0}^{\bar{s}} s dF_\pi(s)}_{=0} - \underbrace{\int_{\check{s}(p,\hat{s})}^{s_0} s\omega_{p,\hat{s}}(s) dF_\pi(s)}_{<0} > 0.$$

Second, suppose $p < \min\{1, \gamma c(\hat{s}|s_0)\}$. Then, setting $p' = \min\{1, \gamma c(\hat{s}|s_0)\}$ is strictly better. Indeed,

$$V_{p',\hat{s}} - V_{p,\hat{s}} = (p' - p) \underbrace{\int_{\check{s}(p',\hat{s})}^{\bar{s}} s dF_\pi(s)}_{\geq 0} + \int_{\check{s}(p',\hat{s})}^{\check{s}(p,\hat{s})} \underbrace{(p - \gamma c(\hat{s}|s))}_{<0} s dF_\pi(s) > 0.$$

Third, suppose $\mu_\pi \geq 0$ and $\gamma c(\hat{s}| - \underline{s}) \leq p < 1$. Then, setting $p' = 1$ (strictly if $\mu_\pi > 0$) is strictly better. Indeed,

$$V_{p',\hat{s}} - V_{p,\hat{s}} = (p' - p) \int_{-\underline{s}}^{\bar{s}} s dF_\pi(s) = (p' - p)\mu_\pi.$$

Finally, suppose $\gamma c(\hat{s}|0) < p$. Then, setting $\hat{s}' = \max\{s \leq \bar{s} : \gamma c(s|0) \leq p\}$ is strictly better. Indeed, this strictly lowers the approval probability of noncompliant states above $\check{s}(p, \hat{s})$ while keeping the approval probability of compliant states constant at p. □

Proof of Theorem 2.

Step 1: $\mathcal{H}^*_{\underline{c}}$ solves (IDE). (IDE) is a linear differential equation with a well-known unique solution:

$$\mathcal{H}(x) = \left\{ \kappa \underline{s} \left(1 + \underbrace{\int_0^x \frac{1}{(\underline{s}+y)\zeta(y)} dy}_{\chi(x)} \right) - \theta \underline{c} \underbrace{\int_0^x \frac{y}{(y-\mu_\pi)(y+\underline{s})\zeta(y)} dy}_{\xi(x)} \right\} \zeta(x),$$

where

$$\zeta(x) = \exp\left(\int_0^x \frac{y}{(y-\mu_\pi)(y+\underline{s})} dy \right).$$

A bit of algebra yields our closed-form expression for $\mathcal{H}^*_{\underline{c}}$. First,

$$\log \zeta(x) = \int_0^x \frac{y}{(y-\mu_\pi)(y+\underline{s})} dy = \left[\frac{\mu_\pi}{\mu_\pi + \underline{s}} \log(y-\mu_\pi) + \frac{\underline{s}}{\underline{s}+\mu_\pi} \log(y+\underline{s}) \right]_0^x,$$

leading to $\zeta(x) = \left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}} \left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}$. Next

$$\xi(x) = \left[-\exp\left(-\int_0^y \frac{z}{(z-\mu_\pi)(z+\underline{s})} dz \right) \right]_0^x = 1 - \frac{1}{\zeta(x)}.$$

Finally, using the closed-form for ζ,

$$\begin{aligned}
\chi(x) &= (-\mu_\pi)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}} \underline{s}^{\frac{\underline{s}}{\mu_\pi+\underline{s}}} \int_0^x (y-\mu_\pi)^{-\frac{\mu_\pi}{\mu_\pi+\underline{s}}} (y+\underline{s})^{-\frac{\underline{s}}{\mu_\pi+\underline{s}}-1} dy \\
&= (-\mu_\pi)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}} \underline{s}^{\frac{\underline{s}}{\mu_\pi+\underline{s}}} \left[\frac{1}{\underline{s}} \left(\frac{y-\mu_\pi}{y+\underline{s}} \right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}} \right]_0^x \\
&= \left(\frac{-\mu_\pi}{\underline{s}} \right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}} \left(\frac{x-\mu_\pi}{x+\underline{s}} \right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}} + \frac{\mu_\pi}{\underline{s}}.
\end{aligned}$$

Plugging these expressions back into our expression for $\mathcal{H}(x)$ yields our closed-form expression, and we get $\mathcal{H}^*_{\underline{c}}$ by choosing κ as indicated, yielding the expression. $\kappa^*_{\underline{c}}$ can be written in closed form as in the body of the paper, or in the following form, which will be useful in proofs

$$\kappa^*_{\underline{c}} = \frac{\overline{s}-\mu_\pi}{\underline{s}\big(1+\chi(\overline{s})\big)\zeta(\overline{s})} + \theta \underline{c} \frac{\zeta(\overline{s})-1}{\underline{s}\zeta(\overline{s})\big(1+\chi(\overline{s})\big)} = \kappa^*_0 + \theta \underline{c} \frac{\zeta(\overline{s})-1}{\underline{s}\zeta(\overline{s})\big(1+\chi(\overline{s})\big)}. \tag{1}$$

Step 2: $\mathcal{H}^*_{\underline{c}}$ is a test function. By construction, $\mathcal{H}^*_{\underline{c}}(\underline{s}) = 0$ and $\mathcal{H}^*_{\underline{c}}(\overline{s}) = \overline{s} - \mu_\pi$. Furthermore, we see from its closed-form expression that $\mathcal{H}^*_{\underline{c}}$ is twice continuously

differentiable, with

$$H^*_{\underline{c}}(x) = \kappa^*_{\underline{c}} + \left(\kappa^*_{\underline{c}}(\mu_\pi + \underline{s}) - \theta\underline{c}\right)\frac{x}{(x+\underline{s})(x-\mu_\pi)}\left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}\mathbb{1}_{x>0},$$

and, differentiating once more,

$$h^*_{\underline{c}}(x) = \left(\kappa^*_{\underline{c}}(\mu_\pi + \underline{s}) - \theta\underline{c}\right)\frac{1}{(x+\underline{s})(x-\mu_\pi)}\left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{-\frac{\underline{s}}{\mu_\pi+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{-\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\mathbb{1}_{x>0}\,. \quad (2)$$

This density has the same sign as $\left(\kappa^*_{\underline{c}}(\mu_\pi + \underline{s}) - \theta\underline{c}\right)$ for $x > 0$, implying it is strictly positive because

$$\begin{aligned}\kappa^*_{\underline{c}}(\mu_\pi + \underline{s}) > \theta\underline{c} &\Leftrightarrow \overline{s} - \mu_\pi > \theta\underline{c}\left(1 + \frac{\overline{s}-\mu_\pi}{\underline{s}+\mu_\pi}\right) = \underline{c}(\overline{s}-\mu_\pi)\\ &\Leftrightarrow \underline{c} < 1.\end{aligned}$$

Hence, $\mathcal{H}^*_{\underline{c}}$ is convex and increasing. Therefore, it must lie below the fully informative test function $\mathcal{H}_{FI}$. It remains to show that $\mathcal{H}^*_{\underline{c}}$ also lies above the uninformative test function $\mathcal{H}_{NI}$. Here, we only show this is true when $\underline{c} = 0$. We show in step 3 that, for every $\underline{c} \in (0,1)$, $\mathcal{H}_{FI} \geq \mathcal{H}^*_{\underline{c}} \geq \mathcal{H}^*_0$, which will expand the conclusion to any $\underline{c}$.

For $\underline{c} = 0$, it is sufficient to show that $H^*_0(\overline{s}) \leq 1$ (note that in our notations, it can be strictly below 1, denoting the presence of an atom at $\overline{s}$). To see why, first note that, by (IDE), $H^*_0(\overline{s}) = \frac{\overline{s}}{\overline{s}+\underline{s}} + \kappa^*_0\frac{\underline{s}}{\underline{s}+\overline{s}}$. Hence, to show $H^*_0(\overline{s}) \leq 1$, it is sufficient to show that $\kappa^*_0 \leq 1$. We can use our closed-form solution to write

$$\begin{aligned}\overline{s} - \mu_\pi = \mathcal{H}^*_0(\overline{s}) &= \kappa^*_0(\overline{s}+\underline{s}) - \kappa^*_0(\mu_\pi+\underline{s}) + \kappa^*_0(\mu_\pi+\underline{s})\left(\frac{\overline{s}-\mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}\\ &= \kappa^*_0(\overline{s}-\mu_\pi) + \kappa^*_0(\overline{s}-\mu_\pi)\left(\frac{\underline{s}+\mu_\pi}{-\mu_\pi}\right)\left(\frac{\overline{s}-\mu_\pi}{-\mu_\pi}\right)^{\frac{-\underline{s}}{\mu_\pi+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}\\ &= \kappa^*_0(\overline{s}-\mu_\pi)\Bigg\{1 + \underbrace{\left(\frac{\underline{s}+\mu_\pi}{-\mu_\pi}\right)\left(\frac{\overline{s}-\mu_\pi}{-\mu_\pi}\right)^{\frac{-\underline{s}}{\mu_\pi+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}}_{\geq 0}\Bigg\}\end{aligned}$$

implying the result.

Step 3: Optimality for the receiver. To see why $\mathcal{H}^*_{\underline{c}}$ is optimal, suppose $\mathcal{H}$ is another test function such that $\mathcal{H}(0) > \mathcal{H}^*_{\underline{c}}(0)$. Without loss of generality, we can take this function to be linear below 0, and let κ be its slope below 0. Then, $\kappa > \kappa^*_{\underline{c}}$ as $\kappa\underline{s} = \mathcal{H}(0) > \mathcal{H}^*_{\underline{c}}(0) = \kappa^*_{\underline{c}}\underline{s}$. Let $x' = \min\{x \in [0,\overline{s}] : \mathcal{H}(x) = \mathcal{H}^*_{\underline{c}}(x)\}$ be the smallest crossing point between $\mathcal{H}$ and $\mathcal{H}^*_{\underline{c}}$. It exists as the minimum of a nonempty

$(\mathcal{H}(\overline{s}) = \mathcal{H}^*_{\underline{c}}(\overline{s}))$ and compact (by continuity of $\mathcal{H} - \mathcal{H}^*_{\underline{c}}$) real subset. Then, we must have

$$H_\ell(x') = \lim_{\substack{x \to x' \\ x < x'}} \frac{\mathcal{H}(x') - \mathcal{H}(x)}{x' - x} \leq \lim_{\substack{x \to x' \\ x < x'}} \frac{\mathcal{H}^*_{\underline{c}}(x') - \mathcal{H}^*_{\underline{c}}(x)}{x' - x} = H^*_{\underline{c}}(x').$$

Then,

$$\begin{aligned} H_\ell(x') - \frac{x}{(\underline{s} + x)(x - \mu_\pi)} \mathcal{H}(x') &\leq H^*_{\underline{c}}(x') - \frac{x}{(\underline{s} + x)(x - \mu_\pi)} \mathcal{H}^*_{\underline{c}}(x') \\ &= \frac{\kappa^*_{\underline{c}} \underline{s}}{\underline{s} + x} - \frac{\theta \underline{c} x}{(x - \mu_\pi)(\underline{s} + x)} \\ &< \frac{\kappa \underline{s}}{\underline{s} + x} - \frac{\theta \underline{c} x}{(x - \mu_\pi)(\underline{s} + x)}, \end{aligned}$$

where the equality is due to the fact that $\mathcal{H}^*_{\underline{c}}$ satisfies (IDE). However, this inequality implies $\mathcal{H}$ does not satisfy (FPIC'). □

Proof of Proposition 9.

Step 1: *We show that for each $\varepsilon > 0$, $(\hat{\tau}_{p_\varepsilon, \varepsilon}, \phi^*)$ is equilibrium feasible.*

Note first that any upward falsification strategy of noncompliant states leads to devaluating signal $\emptyset$, regardless of how compliant states are falsifying. Therefore, the only falsification strategies of noncompliant states that are possibly beneficial for the agent must target $\overline{s}$. To be beneficial, such a strategy must ensure the posterior mean associated with $\overline{x}$ remains positive. Consider a falsification strategy such that a mass m of noncompliant states falsify as $\overline{s}$, and assume it keeps the posterior mean associated with $\overline{x}$ above 0. Then, consider the alternative falsification strategy that concentrates this mass on the higher noncompliant states, that is where only states in $[z, 0)$ with $m = F_\pi(0) - F_\pi(z)$ falsify as $\overline{s}$. The falsification cost of this alternative strategy must be lower by cost monotonicity. Furthermore, the posterior mean following $\overline{x}$ must increase, because falsification originates from higher noncompliant states, so both strategies lead to the same ex-ante approval probability. Hence, the alternative falsification strategy dominates the former, implying we can restrict attention to falsification strategies such that the mass of falsifying noncompliant states is concentrated on an interval $[z, 0)$.

Suppose no noncompliant state falsifies as $\overline{s}$. Then, falsifying any mass of states in $[0, \overline{s})$ as $\overline{s}$ leads to a gain equal to $1 - p - \gamma c(\overline{s}|s)$ for each state, which is positive given our assumption that $p < 1 - \gamma c(\overline{s}|0)$. Indeed, the posterior mean following $\emptyset$ is 0, whereas the posterior mean associated with $\overline{x}$ remains positive. Therefore, falsifying all states in $[0, \overline{s})$ as $\overline{s}$ is optimal.

If an interval $[z, 0)$ of noncompliant states falsifying as $\overline{s}$ exists, signal $\emptyset$ is devaluated, so states in $[0, \overline{s})$ are rejected unless they also falsify as $\overline{s}$. Because the latter can only

increase the posterior mean associated with $\overline{x}$, the gain for each compliant state falsified to $\overline{s}$ is $1 - p - \gamma c(\overline{s}|s) > 0$. Therefore, falsifying all states in $[0, \overline{s})$ as $\overline{s}$ is optimal.

Overall, these arguments imply we can restrict attention to the family of falsification strategies

$$\phi_z(s) = \begin{cases} \delta_{\overline{s}} & \text{if } s \geq z \\ \delta_s & \text{otherwise,} \end{cases},$$

with $z \leq 0$. Note $\phi^* = \phi_0$. Whenever $z < 0$, signal $\emptyset$ is devaluated. If $z < s_0$, signal $\overline{x}$ is also devaluated and rejection is certain, so we can restrict attention to $z \geq s_0$. The agent's payoff is then

$$\int_z^{\overline{s}} \{1 - p - c(\overline{s}|s)\} dF_\pi(s) - r_\varepsilon(p)\big[F_\pi(0) - F_\pi(s_0 - \varepsilon)\big],$$

which is maximized by choosing z equal to $z(p) = \min\{s \geq s_0 : \gamma c(\overline{s}|s) \leq 1 - p\}$. Overall, we have shown the best deviation from ϕ^* is $\phi_{z(p)}$. The net gain of the agent if she deviates to $\phi_{z(p)}$ is

$$\begin{aligned} \Gamma(p, \varepsilon) &= \int_{s_0}^0 \big\{1 - p - \gamma c(\overline{s}|s)\big\}^+ dF_\pi(s) - p\big[1 - F_\pi(0)\big] - r_\varepsilon(p)\big[F_\pi(0) - F_\pi(s_0 - \varepsilon)\big] \\ &= \int_{s_0}^0 \big\{1 - p - \gamma c(\overline{s}|s)\big\}^+ dF_\pi(s) - p\big[1 - F_\pi(0)\big] + p\frac{\int_0^{\overline{s}} s dF_\pi(s)}{\mathbb{E}_\pi\big(s|0 \geq s \geq s_0 - \varepsilon\big)}. \end{aligned}$$

This function is decreasing and continuous in p and $-\varepsilon$. Furthermore, $\Gamma(0, \varepsilon) > 0 > \Gamma(1 - \gamma c(\overline{s}|0), \varepsilon)$, so a unique value $p_\varepsilon \in \big(0, 1 - \gamma c(\overline{s}|0)\big)$ exists such that $\Gamma(p_\varepsilon, \varepsilon) = 0$, and to ensure $\phi_{z(p)}$ is not a profitable deviation, we must therefore choose $p \geq p_\varepsilon$. Hence, $\hat{\omega}_{p,\varepsilon}$ is equilibrium feasible for every $p \geq p_\varepsilon$.

Step 2: *The receiver's payoff is decreasing in ε.*

The receiver's payoff from the equilibrium outcome $\hat{\omega}_{p,\varepsilon}$ is

$$\int_0^{\overline{s}} s dF_\pi(s) + r_\varepsilon(p) \int_{s_0 - \varepsilon}^0 s dF_\pi(s) = (1 - p) \int_0^{\overline{s}} s dF_\pi(s),$$

which is decreasing in p. Hence for any $\varepsilon > 0$, the best equilibrium outcome is $\hat{\omega}_{p_\varepsilon,\varepsilon}$. Furthermore, p_ε is increasing in ε, so the receiver's payoff at the equilibrium outcome $\hat{\omega}_{p_\varepsilon,\varepsilon}$ is also decreasing in ε. As $\varepsilon \to 0$, p_ε converges to $p_0 \leq p_\varepsilon$, where p_0 is the unique value of p such that $\Gamma(p, 0) = 0$, and the receiver's payoff converges to $\hat{V} = (1 - p_0) \int_0^{\overline{s}} s dF_\pi(s)$.

Step 3: *The limit payoff of the receiver is strictly higher than V_γ^* in the low-cost region.*

The value of p_0 is characterized by the formula

$$\Gamma(p_0, 0) = \int_{z(p_0)}^{0} \big\{1 - p_0 - \gamma c(\overline{s}|s)\big\} dF_\pi(s) - p_0\big[1 - F_\pi(s_0)\big] = 0,$$

implying

$$p_0 = \frac{\int_{z(p_0)}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\} dF_\pi(s)}{1 - F_\pi(s_0) + F_\pi(0) - F_\pi(z(p_0))} < \frac{\int_{s_0}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\} dF_\pi(s)}{F_\pi(0) - F_\pi(s_0)}, \tag{3}$$

because, in the low-cost region, we have $\gamma c(\overline{s}|s) \leq \gamma c(\overline{s}|s_0) \leq 1$, for all $s \in [s_0, 0]$. The difference between the limit receiver payoff $\hat{V}$ and V_γ^* in the low-cost region is

$$\begin{aligned}
\hat{V} - V_\gamma^* &= (1 - p_0) \int_0^{\overline{s}} s dF_\pi(s) - \left(\gamma c(\overline{s}|s_0) \int_{s_0}^{\overline{s}} s dF_\pi(s) - \int_{s_0}^{0} \gamma c(\overline{s}|s) s dF_\pi(s)\right) \\
&= (1 - p_0) \int_{s_0}^{0} (-s) dF_\pi(s) - \int_{s_0}^{0} \gamma c(\overline{s}|s)(-s) dF_\pi(s) \\
&= \int_{s_0}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\}(-s) dF_\pi(s) - p_0 \int_{s_0}^{0} (-s) dF_\pi(s) \\
&> \int_{s_0}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\}(-s) dF_\pi(s) - \frac{\int_{s_0}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\} dF_\pi(s)}{F_\pi(0) - F_\pi(s_0)} \int_{s_0}^{0} (-s) dF_\pi(s) \\
&\geq 0,
\end{aligned}$$

where the first inequality is from (3), and we repeatedly use that $\int_{s_0}^{\overline{s}} s dF_\pi(s) = \int_{s_0}^{0} s dF_\pi(s) + \int_0^{\overline{s}} s dF_\pi(s) = 0$. For the last inequality, consider the two probability distributions on $[s_0, 0]$ defined by

$$dG(s) = \frac{dF_\pi(s)}{F_\pi(0) - F_\pi(s_0)}, \qquad \text{and} \qquad dH(s) = \frac{\big(1 - \gamma c(\overline{s}|s)\big) dF_\pi(s)}{\int_{s_0}^{0} \big\{1 - \gamma c(\overline{s}|s)\big\} dF_\pi(s)}.$$

The ratio $\frac{dG(s)}{dH(s)}$ is proportional to $1/\big(1 - \gamma c(\overline{s}|s)\big)$, which is increasing in s, implying G dominates H in the likelihood ratio order, so $\mathbb{E}_G(s) \geq \mathbb{E}_H(s)$, which yields the desired inequality. □

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," *Econometrica*, 87, 1115–1153.

Aumann, R. J. (1974): "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 67–96.

Ball, I. (2020): "Scoring Strategic Agents," *Working Paper*.

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

Bergemann, D. and S. Morris (2016): "Bayes correlated equilibrium and the comparison of information structures in games," *Theoretical Economics*, 11, 487–522.

——— (2019): "Information design: A unified perspective," *Journal of Economic Literature*, 57, 44–95.

Bizzotto, J., J. Rüdiger, and A. Vigier (2020): "Testing, disclosure and approval," *Journal of Economic Theory*, 187, 105002.

Blackwell, D. (1951): "The Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, University of California Press, Berkeley, 93–102.

——— (1953): "Equivalent Comparisons of Experiments," *Annals of Mathematical Statistics*, 24, 265–272.

Cunningham, T. and I. Moreno de Barreda (2015): "Equilibrium Persuasion," *Working Paper*.

Deneckere, R. and S. Severinov (2017): "Screening, Signalling and Costly Misrepresentation," *Working Paper*.

Forges, F. (1986): "An Approach to Communication Equilibria," *Econometrica*, 54, 1375–1385.

Frankel, A. and N. Kartik (2019): "Muddled information," *Journal of Political Economy*, 127, 1739–1776.

——— (2021): "Improving information from manipulable data," *Working Paper*.

Gentzkow, M. and E. Kamenica (2016): "A Rothschild-Stiglitz Approach to Bayesian Persuasion," *American Economic Review: Papers and Proceedings*, 106, 597–601.

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

Kamenica, E. (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11, 249–272.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Kartik, N. (2009): "Strategic communication with lying costs," *Review of Economic Studies*, 76, 1359–1395.

Kartik, N., M. Ottaviani, and F. Squintani (2007): "Credulity, Lies, and Costly Talk," *Journal of Economic Theory*, 134, 93–116.

Kattwinkel, D. (2019): "Allocation with Correlated Information: Too good to be true," *Working Paper*.

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

Kolotilin, A. (2018): "Optimal information disclosure: A linear programming approach," *Theoretical Economics*, 13, 607–635.

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

Lipnowski, E., D. Ravid, and D. Shishkin (2019): "Persuasion via weak institutions," *Working Paper*.

Myerson, R. B. (1982): "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics*, 10, 67–81.

Nguyen, A. and T. Y. Tan (2020): "Bayesian Persuasion with Costly Messages," *Available at SSRN 3298275*.

Ortner, J. and S. Chassang (2018): "Making corruption harder: Asymmetric information, collusion, and crime," *Journal of Political Economy*, 126, 2108–2133.

Roesler, A.-K. and B. Szentes (2017): "Buyer-optimal learning and monopoly pricing," *American Economic Review*, 107, 2072–80.

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

Severinov, S. and T. Y.-C. Tam (2019): "Screening Under Fixed Cost of Misrepresentation," *Working Paper*.

Sobel, J. (2020): "Lying and deception in games," *Journal of Political Economy*, 128, 907–947.

Spence, A. M. (1973): "Job Market Signaling," *Quarterly Journal of Economics*, 87, 355–374.

Terstiege, S. and C. Wasser (2020): "Buyer-optimal extensionproof information," *Journal of Economic Theory*, 105070.