DISCUSSION PAPER SERIES

DP15627

TEST DESIGN UNDER FALSIFICATION

Vasiliki Skreta and Eduardo Perez-Richet

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

TEST DESIGN UNDER FALSIFICATION

Vasiliki Skreta and Eduardo Perez-Richet

Discussion Paper DP15627
Published 03 January 2021
Submitted 24 December 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

TEST DESIGN UNDER FALSIFICATION

Abstract

We study the optimal design of tests with manipulable inputs: data, actions, reports. An agent can, at a cost, falsify the input into the test, or state of the world, so as to influence the downstream binary decision of a receiver informed by the test. We characterize receiver-optimal tests under different constraints. Under covert falsification, the receiver-optimal test is inefficient. With a rich state space, it involves equilibrium falsification at a possibly large cost to the agent, and may therefore exert a negative social externality. The receiver-optimal test that is immune to falsification, while also inefficient, strictly improves the payoff of the agent. When the falsification strategy of the agent is observable, or can be committed to, the receiver-optimal test is efficient, uses a rich signal space, and gives the receiver at least half of his full information payoff.

JEL Classification: C72, D82

Keywords: information design, Falsification, Tests, manipulation, Cheating, Bayesian persuasion

Vasiliki Skreta - vskreta@gmail.com
UT Austin and UVL and CEPR

Eduardo Perez-Richet - edurichet@gmail.com
Sciences Po and CEPR

Test Design under Falsification[*]

Eduardo Perez-Richet[†] Vasiliki Skreta [‡]

December 23, 2020

Abstract

We study the optimal design of tests with manipulable inputs: data, actions, reports. An agent can, at a cost, falsify the input into the test, or state of the world, so as to influence the downstream binary decision of a receiver informed by the test. We characterize receiver-optimal tests under different constraints. Under covert falsification, the receiver-optimal test is inefficient. With a rich state space, it involves equilibrium falsification at a possibly large cost to the agent, and may therefore exert a negative social externality. The receiver-optimal test that is immune to falsification, while also inefficient, strictly improves the payoff of the agent. When the falsification strategy of the agent is observable, or can be committed to, the receiver-optimal test is efficient, uses a rich signal space, and gives the receiver at least half of his full information payoff.

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

1 Introduction

Important decisions are increasingly guided by tests, ratings and algorithms. Criminal justice relies on algorithms to assess risks (Stevenson and Doleac, 2019), employers use them for hiring decisions (Brynjolfsson and Mitchell, 2017). Cars are

[*] We thank the Editor and 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 (ANR STRATCOM - 16-TERC-0010-01). Vasiliki Skreta acknowledges funding by the European Research Council (ERC) consolidator grant 682417 "Frontiers In Design." Angelos Diamantopoulos, Alkis Georgiadis-Harris, 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

tested for emissions. Firms advertise and price their products on the basis of consumer data processed by third-party algorithms. Financial assets are rated, and financial institutions stress tested. Teachers prepare their students to pass tests in order to gain admission to selective schools and universities. This list is suggestive of how wide-ranging and relevant these information processing technologies are, and why it is important that their results are reliable: Fairness, inadequacy, financial distraught, and environmental pollution are at stake when they are compromised. However, their inputs can be, and often are, manipulated.[1]

This paper studies optimal test design with input manipulation, in the form of costly state falsification. Stress tests are run on asset portfolios reported by banks that may hide assets from their balance sheet. Car manufacturers have infamously compromised emissions tests through the use of "defeat devices" that would alter emissions levels in testing conditions. Consumers can alter their behavior to manipulate pricing and advertising algorithms. Teachers can teach their students to the test. All these manipulations amount to falsifying the state of the world on which the test is producing information.

We consider a designer-agent-receiver model of information production. The receiver wishes to take a binary approve-reject decision based on the hidden type of the agent, or state. The designer commits to a test (a Blackwell experiment) that takes the state as an input, and outputs an informative signal for the use of the receiver. Knowing the test, the agent can falsify his type before it is mapped to a signal, so as to maximize approval probability. The resulting equilibrium information structure is co-determined by the test and the falsification strategy of the agent.

How and to what extent falsification impacts equilibrium information structures depends on whether or not it is observable, and on the nature of falsification costs. The fact that costs matter is simple to grasp: if falsification is prohibitively costly, then its impact is null. What is perhaps less obvious is the role of observability. Under covert falsification, the receiver's posterior given a signal is based on anticipated rather than actual falsification. If the agent can commit to his falsification strategy, for example because it is directly observable by the receiver, or detectable, the receiver's posterior reacts to deviations, and the ensuing action

[1] Among prominent examples is the "Diesel emissions scandal" which arose when several manufacturers were found to be cheating on pollution emission tests. Baruchson-Arbib and Bar-Ilan (2007) document search engine manipulation and Hu, Immorlica, and Vaughan (2019) study algorithmic manipulations.

may switch from approval to rejection or vice versa. When falsification is covert, the agent's problem is a signaling, or costly misreporting problem. When falsification is observable, the agent has commitment and he acts as a constrained information designer who can only induce information structures that are feasible given the test. This conceptual difference makes the formal analysis of the two cases very different.

We start with covert and costly falsification. Under a binary state space, we prove a falsification-proofness principle (which also holds for observable falsification) according to which any feasible receiver payoff can be attained with a test that is immune to falsification. We characterize the entire feasible payoff space under falsification-proof tests, and show that the receiver-optimal test is inefficient unless falsification costs are sufficiently high to make all falsification dominated. With continuous one-dimensional state space, the falsification-proofness principle no longer holds. Theorem 1 characterizes a receiver-optimal test under the assumptions that lower states find it harder to falsify as the highest than the lowest state (*costlier to top*), and that the marginal cost of increased upward falsification is higher for higher states (*upward increasing differences*). This test involves equilibrium falsification by the agent, and is therefore inefficient. When falsification costs are low, it also leads to inefficient approval patterns, but attains the receiver's first best for higher costs, and converges to the fully informative outcome as costs rise. The optimal test design problem is in this case equivalent to an optimal mechanism design problem with a single good to allocate, no transfers, and costly misreporting.

Next, motivated by the possibly heavy falsification cost that burdens the agent under the receiver-optimal test, as well as the idea that cheating and lying may exert negative social externalities,[2] we study the optimal design of falsification-proof tests. Theorem 2 characterizes the receiver-optimal falsification-proof test for cost functions that satisfy either increasing or decreasing differences for upward falsification. For decreasing differences, we can use a first-order approach. Under increasing differences, our characterization operates by building an auxiliary program that is the dual of a classical optimal transport problem with well-known solutions. We show that this test is also inefficient, but is indeed strictly better for the agent than the receiver-optimal test of Theorem 1. We also extend our

[2] As documented for other forms of cheating in, for example, Galbiati and Zanella (2012); Ajzenman (2018); Alm, Bloomquist, and McKee (2017); Rincke and Traxler (2011).

method to characterize all constrained efficient tests under falsification-proofness.

Finally, Theorem 3 characterizes a receiver-optimal test under observable falsification. For this more complicated problem, we focus on the binary-state case in which the falsification-proofness principle holds. The optimal test exploits the following trade-off: while upward falsification may lead to better grades, it devalues their meaning when it is observable. This creates an endogenous cost of upward falsification which can be leveraged to deliver better information even when the intrinsic cost of upward falsification is zero. We first characterize our optimal test assuming that only upward falsification is available to the agent, and then extend our result to any environment in which a certain linear combination of upward and downward falsification cost is sufficiently high. In a stark contrast with the covert case, the receiver-optimal test is always efficient, and it gives the receiver at least half of her full-information payoff. To maximally exploit the devaluation channel, the optimal test generates a single rejected signal, and a continuum of approved signals with varying associated beliefs. The more the agent falsifies upward, the higher the fraction of these signals are devalued, making the agent exactly indifferent between truth-telling and any level of upward falsification. We finish the paper by characterizing a subset and a superset of all feasible payoffs under falsification-proof tests and observable falsification, and comparing it to the covert case.

Related Literature. The information design literature[3] has initially focused on the problem of shaping decisions at the receiving end of the the informational process as in Kamenica and Gentzkow (2011) or Bergemann and Morris (2016). By introducing agency in the form of costly state falsification, we contribute to a growing literature that studies information design with moral hazard, or agency: participation in Rosar (2017), effort in Boleslavsky and Kim (2018), Rodina (2016), and Rodina and Farragut (2016), pricing in Roesler and Szentes (2017), additional disclosure in Bizzotto, Rüdiger, and Vigier (2020) and Terstiege and Wasser (2020), signal manipulation in Lipnowski, Ravid, and Shishkin (2019) and Nguyen and Tan (2020), attention in Lipnowski, Mathevet, and Wei (2020) and Bloedel and Segal (2020). In this literature, the choice of information structure also shapes the decisions of an "agent"[4] who can interfere with the information production process

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

[4] In some of these papers, the receiver or the sender acting at different stages of the process.

upstream of the receiver.[5] Within this literature, Frankel and Kartik (2019, 2020) and Ball (2020) essentially study the same agency friction as us, in a setting where the receiver has a continuum of actions, and seeks to most accurately match the agent's fundamental type, which is multi-dimensional in Ball (2020). In addition to the fundamental type, the agent has a privately known gaming ability. These papers study the design of linear scores taking the agent's action as an input. Heterogeneous gaming ability across identical fundamental types is instrumental for manipulations to lead to information muddling in their framework. In our setup, manipulations are effective even without heterogeneity in gaming abilities. Furthemore, we characterize optimal tests without restrictions on the class the designer can choose from. Cunningham and Moreno de Barreda (2015) model test manipulation as state falsification and study equilibrium under a fixed testing technology.

Our work is also related to the contracting and mechanism design literatures with costly state falsification, or misreporting costs. Lacker and Weinberg (1989) incorporate costly state falsification in a model of risk sharing contracts. They characterize optimal falsification-proof contracts, but also show, by example, that optimal contracts may require falsification, without characterizing an optimal contract with falsifiction. In contrast, we characterize an optimal test, which can also be interpreted as an optimal allocation rule, in Theorem 1, as well as a falsification-proof optimal rule in Theorem 2. Landier and Plantin (2016) study optimal taxation with agents that can evade or avoid taxes by concealing income. The literature on mechanism design with reporting costs (Kephart and Conitzer, 2016; Deneckere and Severinov, 2017; Severinov and Tam, 2019), which stems both from economics and computer science, has focused on mechanisms with transfers. In all these papers, except Lacker and Weinberg (1989), the authors make assumptions on the cost function to ensure that a falsification-proofness principle is at work.

Under covert falsification, we show that the problem of finding a receiver-optimal test is formally equivalent to an optimal allocation problem without transfers. To solve a similar problem, the designer can exploit costly verification in Ben-Porath, Dekel, and Lipman (2014), and private information correlated with the agent's type in Kattwinkel (2019), whereas, in our case, optimal design relies

[5] In this sense, we relate to Du (2018) who finds an optimal mechanism under the worst case information structure, while our receiver optimal tests aim to be robust to endogenous agent-driven falsification.

on costly misreporting.

Our test designer plays a role similar to that of a mediator in the mediation literature (Aumann, 1974; Myerson, 1982, 1986, 1991; Forges, 1986). In the covert falsification case, we mainly differ by introducing costly misreporting to the mediation problem. In the observable case, we also introduce the possibility for the reporting agent to commit to his reporting strategy.

Costly falsification can be interpreted as arising from lying cost, which connects our paper to the literature on costly lying (Abeler, Nosenzo, and Raymond, 2019; Kartik, Ottaviani, and Squintani, 2007; Kartik, 2009; Gneezy, Kajackaite, and Sobel, 2018; Guo and Shmaya, 2020; Sobel, 2020). It can also be thought of as signaling (Spence, 1973), where each type of the agent has a *natural* (least costly) action, and the test takes these actions as inputs (browsing behavior, for example, to provide information about preferences). The agent might be led by the test to choose a different action so as to influence the decision of the receiver. The cost of falsification is then opportunity cost of deviating from the natural action. This is in fact the description adopted in Frankel and Kartik (2019, 2020) and Ball (2020).

2 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. To make her decision the receiver can rely on information provided by a *test* that takes the state of the world as an input and outputs an informative signal. However, the agent can manipulate the test by *falsifying* the state of the world.

States and Payoffs. We normalize the receiver's rejection payoff to 0 and equate the state of the world $s \in S \subseteq [-\underline{s}, \overline{s}]$ with her approval payoff. The agent obtains payoff 1 upon approval, and 0 otherwise. We assume $-\underline{s} < 0 < \overline{s}$, and $\{-\underline{s}, \overline{s}\} \subseteq S$. We will focus on the *binary state* case $S = \{-\underline{s}, \overline{s}\}$ and the *continuous state* case $S = [-\underline{s}, \overline{s}]$.

Prior. The prior distribution for the state of the world has probability measure π, with full support on S. We denote its cdf as F_π, and its mean as $\mu_\pi = \mathbb{E}_\pi(s)$.

We let $S^- = S \cap (-\infty, 0)$ and $S^+ = S \cap [0, \infty)$ denote the sets of negative and nonnegative states. If $\mu_\pi < 0$, we let $s_0 = \max\big\{s' \in S : \mathbb{E}_\pi(s|s \geq s') \leq 0\big\}$ denote the largest state such that the receiver would approve if she knew that all lower states are excluded. In particular, if π has no atom at s_0, then $\mathbb{E}_\pi(s|s \geq s_0) = 0$. For convenience, we adopt the convention that $s_0 = -\underline{s}$ when $\mu_\pi \geq 0$. In the binary state case, slightly abusing notations, we let π denote the probability of the high state $= \pi(\overline{s})$. If $\mu_\pi < 0$, 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. For convenience, we adopt the convention that $\varphi_0 = 1$ 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 (in the absence of falsification). The prior π and the test τ together define a joint probability measure on $X \times S$ that we denote by $\tau\pi$. Conditional on observing x, a receiver forms a belief about S that, in the absence of falsification, is given by the conditional probability measure which we denote by $\tau\pi_x$.

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, then $\phi(T|s)$ 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 induced 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. Falsifying s as t comes at cost $\gamma c(t|s)$, where $\gamma \geq 0$ is a scaling parameter, and $c : [-\underline{s}, \overline{s}]^2 \to \overline{\mathbb{R}}_+$ a function that is measurable on $S \times S$. Together, the prior π and the falsification strategy ϕ define the joint probability measure $\phi\pi$ on $S \times S$. The cost of falsification strategy ϕ is then given by $C(\phi) = \gamma \int_{S \times S} c \, d\phi\pi$. Falsification costs may capture expected fines for being caught; explicit technological fabrication or falsification costs such as the cost of defeat devices for emissions test, or the cost of hiding income from tax authorities as in Landier and Plantin (2016); psychological lying costs;[6]; costs due to reputation losses in case cheating

[6] There is ample empirical evidence that lying is costly as documented in the papers we discuss in the literature review.

is discovered; or the opportunity cost of deviating from one's natural action in the signaling interpretation.

Information Structure. 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. If the state space is binary, falsification induces a garbling of τ, but that is not necessarily the case otherwise. In particular, the receiver may prefer $\tau\phi$ to τ (as illustrated in Example 2 below), and this plays an important role in our results.

Observability. An important distinction is whether falsification is *observable* or *unobservable* to the receiver. As discussed in the introduction, when falsification is observable (or detectable from the empirical distribution of test results), the agent is akin to a constrained information designer and can only induce information structures that are feasible given the (exogenous) test in place and his falsification capabilities. In contrast, covert (henceforth, unobservable) falsification is analogous to costly misreporting, or signaling. We treat both cases.

Timing. The timing of the game is as follows:

1. **Test:** A test τ is exogenously given and publicly observable.

2. **Falsification:** The agent 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 having observed ϕ or not.

Note that the agent chooses his falsification strategy *ex ante*, before the state is realized. While this is important if falsification is observable, we show that ex ante and *interim* falsification (knowing the state) are essentially equivalent in the unobservable case (see Lemma B.2).

Solution Concept and Equilibrium. Our equilibrium concept is perfect Bayesian equilibrium. We say that an information structure (τ, ϕ) is (observable or unobservable) *equilibrium feasible* if: (i) The receiver's posterior is derived using Bayes' rule whenever possible: given $\tau\phi$ in the unobservable case, and given $\tau\phi'$ following any choice ϕ' in the observable case; (ii) The receiver approves whenever her posterior mean payoff is nonnegative; (iii) The agent's falsification strategy ϕ is optimal given the receiver's approval strategy.

Note that existence of an equilibrium is not granted under every test, as illustrated in Example 1.

Posterior Beliefs. For each signal x occurring with positive probability according to $\tau\phi\pi$, 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 (slightly abusing notations), we denote this belief by $\tau\phi\pi_x$. Under unobservable falsification, this posterior belief is unaffected by a deviation of the agent. Let $\mu(x|\tau, \phi) = \int_S s \, d\tau\phi\pi_x$ denote the associated posterior mean.

Receiver-Optimal Actions. Given τ, a receiver anticipating ϕ optimally approves whenever she receives a signal x such that $\mu(x|\tau, \phi) \geq 0$. We denote this approval set by $\bar{X}(\tau, \phi) = \{x : \mu(x|\tau, \phi) \geq 0\}$. The receiver's expected payoff is then given by the posterior mean conditional on approval:

$$V(\tau, \phi) = \int_{\bar{X}(\tau,\phi)\times S} \mu(x|\tau, \phi) d\tau\phi\pi.$$

Equilibrium Falsification. Given τ, the *interim* probability that state s is approved if the receiver anticipates ϕ, but the agent secretly deviates to ϕ', is $a(s, \tau, \phi, \phi') = \int_X \mathbb{1}_{\bar{X}(\tau,\phi)} \, d\tau\phi'(s)$. The corresponding ex ante probability of approval is $A(\tau, \phi, \phi') = \int_S a(s, \tau, \phi', \phi) d\pi$, and the payoff of this deviation for the agent is $A(\tau, \phi, \phi') - C(\phi')$. If the agent's falsification choice is correctly anticipated, we denote his payoff as:

$$U(\tau, \phi) = A(\tau, \phi, \phi) - C(\phi).$$

Under *unobservable* falsification, the pair (τ, ϕ) is equilibrium feasible if and only if, for every falsification strategy ϕ',

$$U(\tau,\phi) \geq A(\tau,\phi,\phi') - C(\phi'). \tag{UEF}$$

If falsification is *observable* and the agent deviates from ϕ to ϕ', the posterior mean associated with each signal realization x becomes $\mu(x|\tau,\phi')$, and the set of approved signals $\bar{X}(\tau,\phi')$. In this case, the pair (τ,ϕ) is (observable) equilibrium feasible if and only if, for every falsification strategy ϕ',

$$U(\tau,\phi) \geq U(\tau,\phi') = A(\tau,\phi',\phi') - C(\phi'). \tag{OEF}$$

The following example illustrates the differences between observable and unobservable falsification and shows how an equilibrium may fail to exist for some tests.

Example 1 (Falsifying a fully informative test)**.** *Suppose that the state space is binary and that $\mu_\pi < 0$. Consider a fully informative test, so that $\tau(-\underline{s})$ and $\tau(\overline{s})$ have disjoint support. Let $\underline{c} = \gamma c(\overline{s}|-\underline{s}) \leq 1$, and $\overline{c} = \gamma c(-\underline{s}|\overline{s}) = \infty$, so the agent never falsifies $\overline{s}$ as $-\underline{s}$, and let $\underline{\phi} = \phi(\overline{s}|-\underline{s})$.*

If falsification is observable, the receiver takes the actual choice of ϕ into account when forming her posterior belief. Following a favorable signal $x \in \operatorname{supp}\tau(\overline{s})$, her expected payoff from approval is $\frac{\pi\overline{s}-(1-\pi)\underline{\phi}\underline{s}}{\pi+(1-\pi)\underline{\phi}}$, so she approves if $\underline{\phi} \leq \varphi_0$. Following a signal $x \in \operatorname{supp}\tau(-\underline{s})$, she is certain that the state is $-\underline{s}$ and rejects. The payoff of the agent is therefore equal to $\big\{\pi + \underline{\phi}(1-\pi)(1-c)\big\}\,\mathbb{1}\big(\underline{\phi} \leq \varphi_0\big)$, so he optimally chooses $\underline{\phi} = \varphi_0$, which is the falsification level that makes the receiver indifferent between both actions when receiving a signal indicating the high state. 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's payoff is the result of concavification. As the falsification cost increases, the agent's payoff falls, but the test and the receiver's payoff remain unchanged.[7]

If, instead, falsification is unobservable, the receiver must first form a belief about ϕ, which must be correct in equilibrium. Because a signal in $\operatorname{supp}\tau(-\underline{s})$

[7] Note that the agent could perform this manipulation on any two-signal test with a signal leading to approval under no falsification. In other words, when falsification is observable, all binary signal tests yield a null payoff to the receiver. This is in stark contrast to the unobservable falsification case where binary tests are canonical (see Lemma B.1).

can only be generated by the low state regardless of ϕ, such a signal must lead the receiver to reject in equilibrium. She may approve after seeing a signal in $\operatorname{supp}\tau(\overline{s})$ *only if the equilibrium choice of the agent satisfies* $\underline{\phi} \leq \varphi_0$*. However, if the equilibrium strategy of the receiver is such that she approves for some signals, the unique best response of the agent is then to choose* $\underline{\phi} = 1$*. Therefore the equilibrium strategy of the receiver must be to always reject. If* $c = 0$*, choosing any* $\underline{\phi} > \varphi_0$ *is a best response of the agent. These choices form an equilibrium in which both players get their worst possible payoff. If* $c > 0$*, however, there is no equilibrium under a fully informative test.* ◇

Receiver-optimal Information Structures. A pair (τ, ϕ) is receiver-optimal if it maximizes $V(\tau, \phi)$ subject to (UEF) if falsification is unobservable, or (OEF) if falsification is observable.

Efficiency Notions. Falsification may create inefficiencies through two channels. The first one is *informational*, as the designer needs to alter the test to prevent harmful falsification. The second one arises from the falsification cost incurred on path by the agent. To distinguish between these sources of inefficiency, we introduce the notion of *informational efficiency*: an equilibrium feasible information structure (τ, ϕ) is informationally efficient if it is efficient gross of falsification costs. That is if, for some $\alpha > 0$, it maximizes $V(\tau', \phi') + \alpha\big[U(\tau', \phi') + C(\phi')\big]$ across all possible information structures (τ', ϕ'). It is immediate to show that:[8]

Lemma 1 (Informational Efficiency)**.** *An information structure is informationally efficient if and only if: (i) when the state is binary, it leads the receiver to approve* $\overline{s}$ *with probability* 1*, and* $-\underline{s}$ *with a probability in* $[0, \varphi_0]$*; (ii) when the state is continuous, there exists a cutoff* $s^\dagger \in [s_0, 0]$ *such that, for almost every state* s*, the interim approval probability is* $a(s) = \mathbb{1}_{s \geq s^\dagger}$*.*

Note that, under all informationally efficient information structures, positive states are approved, and states below s_0 rejected, with certainty. We say that an informationally inefficient information structure features *inefficient rejection* if

[8] An information structure is associated with an interim probability of approval $a(s)$. It is informationally efficient iff it maximizes $\int_{-\underline{s}}^{\overline{s}} a(s)(s + \alpha)d\pi$ subject to receiver obedience, where $\alpha \geq 0$ captures the relative weight on the agent. Pointwise maximization implies that the approval probability optimally switches from 0 to 1 at $-\alpha$, but receiver obedience binds for $-\alpha < s_0$, restricting possible thresholds.

it accepts positive states with probability less than 1, and *inefficient approval* if approval probabilities over negative states differ from those in the lemma.

Falsification Proofness Principle. When possible, it is helpful to rely on a revelation-principle type of result allowing us to restrict attention to tests that induce truth-telling as an equilibrium falsification strategy. We now establish that such a falsification-proofness principle holds when falsification is costless, or when the state space is binary.

Proposition 1 (Falsification Proofness Principle)**.** *If falsification is costless, or the state space is binary, then, for every equilibrium feasible information structure (τ, ϕ), the test $\tau' = \tau\phi$ and the truth-telling strategy δ form an equilibrium feasible pair. Furthermore $V(\tau, \phi) = V(\tau', \delta)$, and $U(\tau', \delta) = U(\tau, \phi) + C(\phi)$. This holds under both observable and unobservable falsification.*

The proof of this result follows a standard line of argument. It is easy in the costless case, as the outcome of any falsification strategy ϕ' under $\tau' = \tau\phi$ can be reached with falsification strategy $\phi\phi'$ under τ. When falsification is costly, however, we need to further argue that $C(\phi'\phi) - C(\phi) \leq C(\phi)$. In the proof of Proposition 1, we show that this is true with a binary state space. The following example shows that the result no longer holds with more than two states.

Example 2. *Suppose that $S = \{-3, 1, 3\}$, the prior is $\pi = \{1/2, 1/4, 1/4\}$, and falsification costs are given by $c(t|s) = |t-s|/5$. Note that falsifying -3 as 3 is never worth it for the agent, as it costs $6/5 > 1$. Consider the deterministic Pass/Fail test τ that maps state 3 to the Pass signal, and other states to the Fail signal. Let ϕ be the strategy falsifying 1 as 3 with probability 1, which is easily seen to be equilibrium feasible under τ, both in the observable and unobservable case. Note that (τ, ϕ) gives the receiver her first-best payoff as all positive states are approved, and all negative states rejected. In particular, the receiver prefers (τ, ϕ) to (τ, δ), illustrating how falsification does not necessarily garble information, and may benefit the receiver. Then, the test $\tau' = \tau\phi$ is one that sends the Pass signal whenever the state is positive, and the Fail signal otherwise. The optimal falsification strategy under τ' is to falsify -3 as 1 with probability 1 in the unobservable case, and with probability $1/2$ in the observable case, implying that truth-telling δ cannot be an equilibrium falsification strategy under τ' in either case.* ◇

Falsification Costs. Without loss of generality, we always define the cost function as a function from $[-\underline{s}, \overline{s}]^2$ to $\overline{\mathbb{R}}_+$. The following are maintained and natural assumptions about the cost function. First, truth-telling is costless, $c(s|s) = 0$. Second, it is monotonic in the sense that is strictly more costly to falsify to further away states.[9] Finally, the cost function is continuous.

The next properties are not always assumed but play an important role for some results. The *costlier-to-top* property says that it is costlier to falsify the threshold state $s = 0$ as the highest state, than as the lowest affordable state. By monotonicity, this property extends to all negative states, and it captures in a relatively unrestrictive manner the intuitive idea that falsifying upward is harder than falsifying downward. The next two properties are opposite statements about the return of falsifying further up as a function of the true state: under (UID), this return is lower for a higher true state, whereas it is higher under (UDD). The last one is a specific smoothness assumption that suits our purpose and that we call *regularity*. When and where they exist, we denote the partial derivatives of the cost function by c_t and c_s.

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) has upward increasing differences if, for every $s < s' \leq t < t'$,

$$c(t'|s') - c(t|s') \geq c(t'|s) - c(t|s); \tag{UID}$$

(iii) has upward decreasing differences if, for every $s < s' \leq t < t'$,

$$c(t'|s') - c(t|s') \leq c(t'|s) - c(t|s); \tag{UDD}$$

(iv) is regular if $c(t|s)$ is continuously differentiable in t on $[s, \overline{s}]$ and in s on $[-\underline{s}, t]$, and there exists $K > 0$ such that, for every $t > s$,

$$c(t|s) \leq K(t - s). \tag{REG}$$

[9] 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 $t \leq s < s'$.

Note that any cost function of the form[10] $c(t|s) = k^{-}(|t-s|)\,\mathbb{1}(t \leq s) + k^{+}(|t-s|)\,\mathbb{1}(t \geq s)$, where k^{+} and k^{-} are nonnegative-valued increasing functions with $k^{-}(0) = k^{+}(0) = 0$ satisfies (UID) if k^{+} is concave, or more generally subadditive;[11] (UDD) if k^{+} is convex, or more generally superadditive; (REG) if k^{+} is continuously differentiable; satisfies (UDD), (UID) and (REG) if k^{+} is linear. (CTT) is the only assumption that bears on downward falsification.

3 Unobservable Falsification

In this section, we study unobservable falsification. In Section 3.1, we establish two key preliminary results: The first one is a *recommendation principle* (without loss we can restrict attention to signal realizations that are action recommendations), and the second one is that an equivalence result between ex-ante and interim optimal falsification. In Section 3.2, we focus on the binary state and, leveraging these two results, as well as the falsification-proofness principle Proposition 1, we easily characterize the receiver-optimal test as well as the entire set of equilibrium feasible information structures. In Section 3.3, we charactrize a receiver-optimal test in the continuous state case. Finally, in Section 3.4, we characterize a receiver-optimal falsification-proof test, and then show how to extend this result to characterize constrained efficient falsification-proof tests.

3.1 Preliminary Results

Recommendation Principle. Mimicking standard results as those in Myerson (1982) and Kamenica and Gentzkow (2011), we establish a recommendation principle according to which any test can be garbled into a binary-signal test whose signal realizations are action recommendations (in our case: approve or reject) without changing equilibrium falsification strategy, payoffs and interim approval probabilities. The garbled test simply pools together all signals leading to the same action, thus ensuring the receiver's obedience while maintaining the same interim approval probabilities. In our setting, however, we also need to make sure

[10] Another example are cost functions of the form $c(t|s) = a(s)k^{+}(|t-s|)\,\mathbb{1}\,(t \geq s) + \frac{1}{b(s)}k^{-}(|t-s|)\,\mathbb{1}\,(t < s)$. In this parameterization, $a(s)$ and $b(s)$ can be interpreted as capturing an agent's gaming ability as in Frankel and Kartik (2019) and Ball (2020), restricted to be perfectly correlated with the state.

[11] In their taxation model, Landier and Plantin (2016) justify such subadditive functions as capturing increasing returns to scale in income hiding, a form of costly state falsification.

that it does not yield a new equilibrium falsification strategy. This part of the proof leverages the fact that, when falsification is unobservable, the set of passing signals does not react to deviations from ϕ. The formal statement (Lemma B.1) and proof of this result are in Online Appendix B. This allows us to restrict our discussion to binary tests such that the receiver obeys recommendations.

For the remainder of our analysis of the unobservable case, we therefore, in a slight abuse of notations, redefine tests as measurable functions $\tau : S \to [0, 1]$, where $\tau(s)$ is the *nominal approval probability* of state s. Falsification may of course induce a *true approval probability* that differs from the nominal one. The receiver obedience constraint is then[12]

$$\int_{S\times S} s\tau(t)d\phi\pi(t,s) \geq \max\{\mu_\pi, 0\}. \tag{RO}$$

Interim–Ex Ante Falsification Equivalence. With this redefinition, the condition characterizing an *ex ante* equilibrium feasible falsification strategy ϕ becomes:

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

If the agent could choose ϕ at the interim stage, after observing the state, the condition for ϕ to be equilibrium feasible would be:

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

It is easy to see[13] that (AEF) is equivalent to the interim conditon holding for almost every s. Because the outcome of falsification from a set of states with measure 0 has no effect on the ex-ante payoffs of the receiver or the agent, we, essentially without loss of generality, restrict attention to falsification strategies that satisfy (IEF).

Costless Falsification. When falsification is costless, the falsification proofness principle of Proposition 1 applies and, combined with (IEF), implies that the test must give a constant passing probability. If $\mu_\pi < 0$, the recommendation principle

[12] The expected state following the approve signal is given by $\int_{S\times S} s\tau(t)d\phi\pi(t,s)$ and must be nonnegative, whereas the expected state following the reject signal is $\int_{S\times S} s(1-\tau(t))d\phi\pi(t,s) = \mu_\pi - \int_{S\times S} s\tau(t)d\phi\pi(t,s)$ and must be nonpositive.

[13] For a formal statement and a proof see Lemma B.2 in Online Appendix B.

implies that the passing probability is 0 for all states, as otherwise all states would be equally likely to generate the passing signal, leading the receiver to expect a payoff of $\mu_\pi < 0$. Similarly, if $\mu_\pi \geq 0$, the recommendation principle implies that the passing probability is 1 for all states.

Corollary 1 (Unobservable and Costless Falsification)**.** *When falsification is unobservable and costless, the unique equilibrium outcome is such that no useful information is provided, and the receiver always rejects if $\mu_\pi < 0$ and always approves if $\mu_\pi \geq 0$.*

Receiver Program. By the recommendation principle and interim–ex ante equivalence, a receiver-optimal test under unobservable falsification can be found by solving the following *receiver program*:

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s) \quad \text{s.t.} \quad \text{(IEF)}, \ \text{(RO)} \tag{RP}$$

Note that the obedience constraint is redundant.[14] Interestingly, this implies that the receiver does not benefit from more commitment power.

Proposition 2. *For the receiver, direct commitment to an approval strategy has no additional value compared to commitment to a test.*

This proposition requires no proof as the program of such a committed receiver is exactly (RP) without the obedience constraint. Note further that this is also the program of a principal seeking to allocate a good to an agent of type s, where s is the value for the principal of allocating the good to the agent; the principal also has an outside option (not allocating the good) worth 0; the agent gets a state independent payoff from getting the good; the principal can commit to a probabilistic allocation rule τ contingent on the reported state; misreporting is costly. As mentioned in the introduction, this connects our analysis to the literature studying allocation problems without transfers.

[14] Indeed, the left-hand side term in the obedience constraint is equal to the expected payoff of the receiver, which is also the objective function of (RP). Since the uninformative test makes falsification irrelevant, therefore satisfying (IEF) for all ϕ, and satisfies the obedience constraint (by choosing $\tau(s) = 1$ if $\mu_\pi \geq 0$ and $\tau(s) = 0$ otherwise), so does the solution to the relaxed problem as it must yield a higher receiver payoff.

3.2 The Binary State Case

Under binary state, combining the falsification-proofness principle and the recommendation principle leads to an almost immediate characterization of the set of constrained efficient payoffs. Indeed, by Proposition 1, it is exactly the set of FP-constrained efficient feasible payoffs.

Adapting our notations to this case, we let $\underline{c} = \gamma c(\overline{s}| - \underline{s})$, and $\overline{c} = \gamma c(-\underline{s}|\overline{s})$. Using the recommendation principle, we denote the test by $\tau = (\underline{\tau}, \overline{\tau})$, where $\underline{\tau}$ is the nominal passing probability of the low state $-\underline{s}$, and $\overline{\tau}$ that of the high state. Then the set of equilibrium feasible approval probabilities is characterized by the obedience constraint

$$\overline{\tau}\pi\overline{s} - \underline{\tau}(1-\pi)\underline{s} \geq \max\{\mu_\pi, 0\}, \tag{RO}$$

and the falsification proofness constraint[15]

$$\overline{\tau} - \underline{\tau} \leq \underline{c}, \tag{FPIC}$$

which define a convex polytope. The sender's payoff $V(\tau, \delta) = \overline{\tau}\pi\overline{s} - \underline{\tau}(1-\pi)\underline{s}$, and the receiver's payoff $U(\tau, \delta) = \pi\overline{\tau} + (1-\pi)\underline{\tau}$ are linear in $(\underline{\tau}, \overline{\tau})$, so the set of equilibrium feasible *payoffs* is also a convex polytope.

Suppose that $\mu_\pi < 0$. Then the uninformative and obedient test $\tau_{NI} = (0,0)$ is pessimal for both players; the fully informative and obedient test $\tau_{FI} = (0,1)$ yields the first best for the receiver, while the agent optimal obedient test is $\tau_{KG} = (\varphi_0, 1)$, where KG stands for Kamenica and Gentzkow (2011) as this is the agentr (aka sender) optimal information structure in their famous example. When $\underline{c} \geq 1$, all these tests also satisfy (FPIC), and the set of equilibrium feasible information structures is $\text{co}\big(\{\tau_{NI}, \tau_{FI}, \tau_{KG}\}\big)$, which coincides with what is feasible without falsification. At the other extreme, when $\underline{c} = 0$ only $\tau_{NI} = (0,0)$ is feasible. We now turn to the interesting range of falsification costs $\underline{c} \in (0,1)$. Elementary algebra yields that $\tau_R = (0, \underline{c})$ is the receiver-optimal test. Coming to the agent, the range $\underline{c} \in (0,1)$ can be divided into two regions depending on whether or not τ_{KG} is feasible. By construction τ_{KG} satisfies (RO) with equality, but it violates (FPIC) when $\underline{c} \leq 1 - \varphi_0$; in this range, the agent-optimal test is the one that satisfies both (RO) and (FPIC) with equality, $\tau_A = \big(-\frac{\underline{c}\pi\overline{s}}{\mu_\pi}, \frac{-\underline{c}(1-\pi)\underline{s}}{\mu_\pi}\big)$. When τ_{KG} satisfies (FPIC) with slack, which happens when $\underline{c} > 1 - \varphi_0$, another extremal

[15] It is easy to show that (RO) implies that the second falsification proofness constraint $\underline{\tau} - \overline{\tau} \leq \overline{c}$ is redundant.

information structure arises: the test $\tau_P = (1 - \underline{c}, 1)$ that satisfies (FPIC) with equality but (RO) with slack.[16] Then, the set of equilibrium feasible tests is:

$$\mathcal{T} = \begin{cases} \tau_{NI} & \text{if } \underline{c} = 0, \\ \operatorname{co}\big(\{\tau_{NI}, \tau_R, \tau_A\}\big) & \text{if } 0 < \underline{c} \leq 1 - \varphi_0, \\ \operatorname{co}\big(\{\tau_{NI}, \tau_R, \tau_{KG}, \tau_P\}\big) & \text{if } 1 - \varphi_0 < \underline{c} < 1, \\ \operatorname{co}\big(\{\tau_{NI}, \tau_{KG}, \tau_{FI}\}\big) & \text{if } \underline{c} \geq 1. \end{cases}$$

We depict $\mathcal{T}$ in Figure 1 for various cost levels. The corresponding set of feasible payoffs is depicted in Figure 7, in Section 4, where we compare it to the set of feasible payoffs under observable falsification.

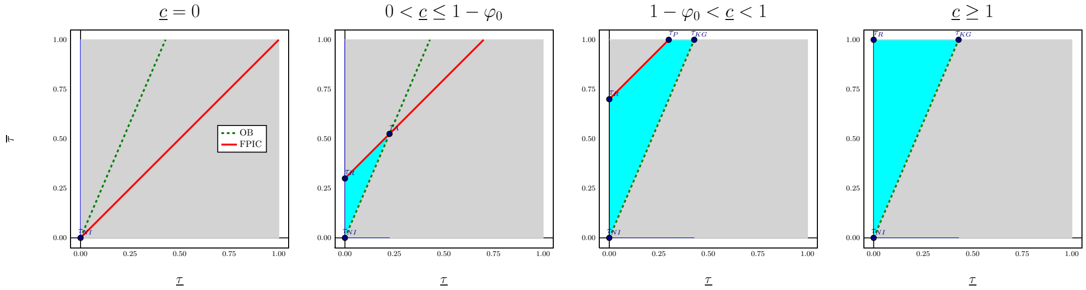

Figure 1: *Unobservable falsification: the blue region is the set of feasible information structures $\mathcal{T}$. Parameters for the plots: $-\underline{s} = -2, \overline{s} = 2, \pi = 0.3, (\mu_\pi = -0.8)$; $\underline{c} \in \{0, 0.3, 0.7, 1\}$.*

With two states we can rely on falsification-proof tests so there is no inefficiency due to incurred costs. However, when $\underline{c} < 1$, the receiver-optimal test is informationally inefficient due to inefficient approval of the high state. Furthermore, if $\underline{c} < 1 - \varphi_0$, there is no efficient (or informationally efficient) feasible test. If $\underline{c} \geq 1 - \varphi_0$, all tests on $\operatorname{co}\big(\{\tau_{KG}, \tau_P\}\big)$ are efficient. As we show next, in the continuous state case, the receiver-optimal test is always inefficient due both to falsification costs incurred by the agent, and also to informational inefficiency for sufficiently low costs.

[16] The test $\tau_P = (1 - \underline{c}, 1)$ is, in fact, the optimal test for a *planner* who assigns equal weights to the receiver and the agent. In all other parameter ranges it coincides with another extremal information structure: it is equal to τ_R when $\mu_\pi < -1$, and, for $-1 \leq \mu_\pi < 0$, it coincides with τ_A in the cost range where τ_{KG} is infeasible.

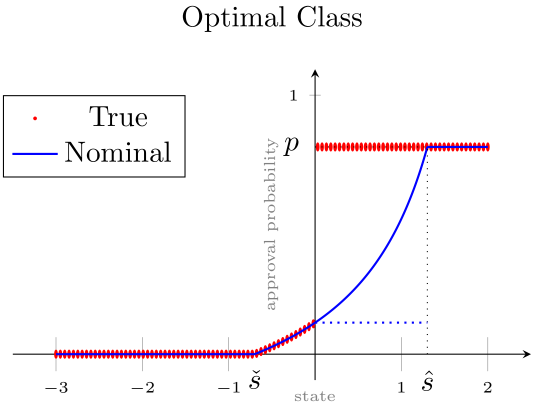

Figure 2: *The blue curve gives the nominal passing probabilities of a test in our class, whereas the red dotted curve shows the true passing probabilities when the agent uses the receiver-preferred equilibrium falsification strategy $\phi_{p,\hat{s}}$. The blue dotted line illustrates alternative nominal passing probabilities for states in $[0, \hat{s}]$ that deliver the same true passing probabilities. The cost function is $c(t|s) = \frac{1.2|t-s|}{1+|t-s|}$ if $t \geq s$.*

3.3 Continuous State Space: A Receiver-Optimal Test

In this section, we characterize a receiver-optimal test under the assumption that the cost function satisfies (UID)[17] and (CTT). For this purpose, we introduce a simple class of tests to which attention can be restricted when looking for a solution of (RP).

An Optimal Class of Tests. Tests in this class are characterized by two parameters: the highest nominal passing probability $p \in [0, 1]$, and the cutoff state, $\hat{s} \in S^+$, above which nominal probabilities are set to p. They are defined as follows:

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

where $\check{s}(p, \hat{s})$ is the cutoff state below which the nominal probability is 0. When it exists, it is implicitly defined as the solution of $\gamma c(\hat{s}|\check{s}) = p$ in $s \in [-\underline{s}, \hat{s}]$. Otherwise, we set $\check{s}(p, \hat{s})$ to $-\underline{s}$.

[17] In the online appendix, all results in this section are proved under the weaker assumption that the cost function satisfies the *upward triangular inequality*: for all $s < m < t$, $c(t|m) + c(m|s) > c(t|s)$. The results are stated with (UID), to facilitate comparisons with results in the next section.

These tests have interesting properties:

Lemma 2. *For every $p \in [0,1]$, and every $\hat{s} \in S^+$, the test $\tau_{p,\hat{s}}$ is strictly increasing on $\big[\check{s}(p,\hat{s}),\hat{s}\big]$ and constant and equal to 0 for $s \leq \check{s}(p,\hat{s})$ and equal to p for $s \geq \hat{s}$. Furthermore:*

(i) If the cost function satisfies (UID), then $s \in \operatorname{argmax}_t \tau_{p,\hat{s}}(t) - \gamma c(t|s)$ for every $s \in S$.

(ii) For every $s \in \big[\check{s}(p,\hat{s}),\hat{s}\big]$, $\hat{s} \in \operatorname{argmax}_t \tau_{p,\hat{s}}(t) - \gamma c(t|s)$.

Part (i) establishes that, under (UID), truth-telling is an optimal choice for the agent for all states. Part (ii) shows that falsifying as $\hat{s}$ is also optimal for all states in the interval $\big[\check{s}(p,\hat{s}),\hat{s}\big]$. This indifference implies that there are multiple optimal falsification strategies for the agent. Among these, the receiver-optimal one is that all positive states in $\big[\check{s}(p,\hat{s}),\hat{s}\big]$ falsify as $\hat{s}$, while all negatives do not falsify. As a result, positive states are approved with probability p, whereas negative states are approved with their nominal approval probability, as illustrated in Figure 2. Let

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

denote this strategy.[18]

Optimizing the receiver's payoff within the class $(\tau_{p,\hat{s}},\phi_{p,\hat{s}})$ reduces the original infinite dimensional problem to a two dimensional one. Theorem 1 characterizes the equilibrium feasible information structure $(\tau_{p,\hat{s}},\phi_{p,\hat{s}})$ that solves the receiver's problem.

Theorem 1. *Suppose that the cost function satisfies (UID) and (CTT). Then the equilibrium feasible information structure $(\tau_{p^*,\hat{s}^*},\phi_{p^*,\hat{s}^*})$ solves (RP), where*

$$\hat{s}^* = \max\big\{s \in S : \gamma c(s|0) \leq 1\big\}, \qquad \text{and } p^* = \begin{cases} \min\big\{\gamma c(\overline{s}|s_0),1\big\} & \text{if } \mu_\pi < 0 \\ 1 & \text{if } \mu_\pi \geq 0 \end{cases}.$$

Furthermore, this equilibrium is always inefficient.

[18] Note that there are other choices of tests that admit $\phi_{p,\hat{s}}$ as an equilibrium falsification strategy, and then lead to the same true approval probabilities. Indeed, any test τ that coincides with $\tau_{p,\hat{s}}$ outside of $[0,\hat{s}]$, and satisfies $\tau(s) \leq \tau_{p,\hat{s}}(s)$ otherwise achieves this. Furthermore, type $s \in [0,\hat{s}]$ has a strict incentive to falsify as $\hat{s}$ under τ.

The crucial step to prove Theorem 1 consists in showing that, when looking for a receiver-optimal test, it is possible to restrict attention to our class. We do this in two steps:

Step 1: Symmetrization. Starting from any equilibrium feasible (τ, ϕ), we construct a new test with an equilibrium such that nonnegative states do not falsify as negative states, and under which both the agent and the receiver are better off. This new test is constructed by *symmetrizing* the original test, that is by replacing the nominal passing probability of any nonnegative state s by $\max\{\tau(s), \tau(\sigma(s))\}$, where $\sigma(s)$ is the negative state such that $c(s|0) = c(\sigma(s)|0)$. The existence of $\sigma(s)$ is ensured by (CTT). By doing so, we provide nonnegative states with better falsification opportunities on the positive side, without giving negative states any new falsification opportunity. This implies that nonnegative states obtain higher true passing probabilities, while nothing changes for negative states, which is good for both the receiver and the agent.

Step 2: Optimality of class. For every equilibrium feasible (τ, ϕ) such that nonnegative states do not falsify as negative states, we show that there exists a test $\tau_{p,\hat{s}}$ from our class, such that the receiver prefers $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$ to (τ, ϕ). To do this, we set[19] $p = \max_{s \geq 0} \tau(s)$, and $\hat{s}$ to be the lowest positive state that satisfies either $\tau(s) = p$ or $\gamma c(s|0) \geq p$. Then it is easy to show that the new test gives every negative state a lower true approval probability. It also gives any nonnegative state approval probability p, which is higher than their true approval probability under (τ, ϕ) since nonnegative states only falsified as nonnegative states.

Properties of Optimal Test. The receiver-optimal test is falsification proof.[20] The receiver and agent payoffs at the receiver optimal information structure are respectively given by

$$V(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*}) = \int_{-\underline{s}}^{0} s\tau_{p^*,\hat{s}^*} dF_\pi(s) + p^* \int_{0}^{\overline{s}} s dF_\pi(s),$$

and

$$U(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*}) = U(\tau_{p^*,\hat{s}^*}, \delta) = \int_S \tau_{p^*,\hat{s}^*}(s) dF_\pi(s).$$

[19] For the sake of providing intuition, we are assuming here that τ is continuous.

[20] In a result available upon request, we have proved that, if the cost function satisfies the triangular, any test can be made falsification-proof for *negative states* while improving the payoffs of both the receiver and the agent. This is related to Kephart and Conitzer (2016), who show that a revelation principle holds under the same condition in mechanism design problems with transfers.

The expression for the agent reflects the fact that he is indifferent between truth-telling and $\phi_{p^*,\hat{s}^*}$, so his payoff can be evaluated as if he was not falsifying.

When falsification is impossible, receiver-optimality, which dictates to approve all states above 0 with probability 1 and to reject all negatives, implies informational efficiency. Under covert falsification, the receiver-optimal equilibrium information structure is inefficient. One source of inefficiency stems from the falsification costs incurred by the agent when the state is between 0 and $\hat{s}^*$. Abstracting from these costs, however, it can also exhibit informational inefficiencies. The nature and degree of inefficiency of $(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*})$ varies with the level of falsification cost, which we capture with the parameter γ.

I. Inefficient rejection and approval: When $\mu_\pi < 0$ and $\gamma c(\overline{s}|s_0) < 1$, then $p^* = \gamma c(\overline{s}|s_0)$ and the true approval probability function is:

$$a^*(s) = \begin{cases} 0 & \text{if } s \leq s_0 \\ \gamma\big\{c(\overline{s}|s_0) - \gamma c(\overline{s}|s)\big\} & \text{if } s_0 < s < 0 \\ \gamma c(\overline{s}|s_0) & \text{if } s \geq 0 \end{cases}. \tag{I}$$

By comparing (I) with the efficient approval probability, it can be readily verified that, in this parameter region, the optimal test exhibits inefficient rejection and approval. To see that the agent's payoff is increasing in γ, note that because the agent is indifferent between falsifying to $\hat{s}^*$ and not falsifying, his payoff is equal to $\gamma\big\{c(\overline{s}|s_0) - c(\overline{s}|s)\big\}$ for all $s \geq s_0$ and zero otherwise, which is increasing in γ because $c(\overline{s}|s_0) - c(\overline{s}|s) \geq 0$. To see that the receiver's payoff is increasing in γ, note that holding p^* fixed as γ increases to γ', we have

$$V^*(\gamma) = V(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*}) \leq \int_{-\underline{s}}^{0} s(p^* - \gamma' c(\overline{s}|s))dF_\pi(s) + p^* \int_0^{\overline{s}} s dF_\pi(s).$$

At γ', $p^*(\gamma')$ and $\hat{s}^*(\gamma')$ are also optimally chosen implying that $V^*(\gamma) \leq V^*(\gamma')$.

II. Inefficient approval: When $\gamma c(\overline{s}|0) < 1$, and, if $\mu_\pi < 0$, $\gamma c(\overline{s}|s_0) \geq 1$, the true approval probability function is:

$$a^*(s) = \begin{cases} \big\{1 - \gamma c(\overline{s}|s)\big\}^+ & \text{if } s < 0 \\ 1 & \text{if } s \geq 0 \end{cases}. \tag{II}$$

Again, one can easily see from (II) that, in this region, there is inefficient approval. It is immediate that the receiver's payoff is increasing in γ (since the probability that negative states are approved decreases as γ increases) while agent's payoff is decreasing in γ because again, the agent's payoff at each s is equal to $\{1 - \gamma c(\bar{s}|s)\}^+$.

III. Receiver first-best. When $\gamma c(\bar{s}|0) \geq 1$, the true approval probability function is :

$$a^*(s) = \begin{cases} 0 & \text{if } s < 0 \\ 1 & \text{if } s \geq 0 \end{cases} . \tag{III}$$

The receiver's payoff is constant and equal to her first best payoff and the outcome is informationally efficient, whereas the agent's payoff eventually becomes increasing in γ. This is because in this region γ affects the agent through a new channel: $\hat{s}^*(\gamma)$ decreases as γ increases: As $\gamma \to \infty$, the threshold $\hat{s}^*(\gamma)$ converges to 0, and thus the range of positive states falsifying vanishes and the agent's payoff is eventually equal to that arising at a fully informative test without falsification. Note that in this region our test is optimal even if the cost function fails both (UID) and (CTT).

The only channel through which the prior distribution affects the optimal test is through parameter s_0 (below which states are never approved), and it is only the case if $\mu_\pi < 0$, and falsification costs are sufficiently low (i.e. in region I).

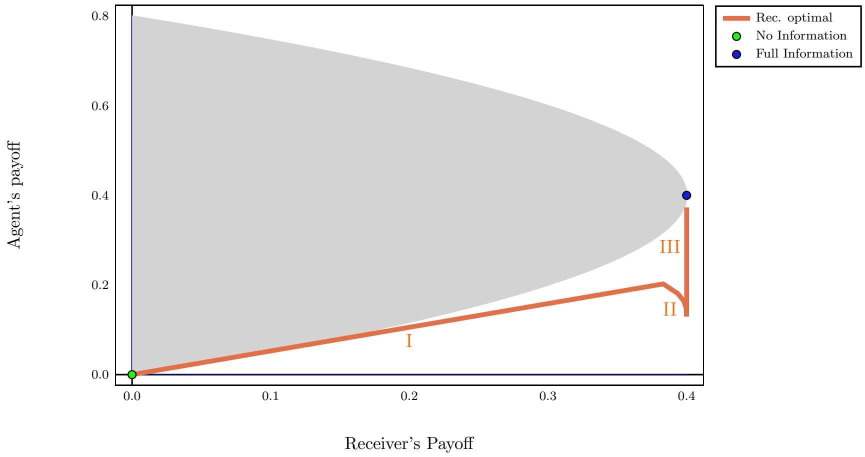

Figure 3: *The grey area depicts the set of attainable payoffs under all possible information structures and no falsification. The orange path shows the payoffs from the receiver-optimal test as a function of γ. The curve starts at the no information payoffs for $\gamma = 0$, moves successively across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

Figure 3 illustrates how the payoff vector at the receiver-optimal test varies with γ (orange curve), as well the set of feasible payoffs in the absence of falsification (in grey). It illustrates the inefficiency of the receiver-optimal test, and the possibly heavy cost to the agent.

3.4 Continuous State Space: Optimal Falsification-Proof Tests

The receiver-optimal test we derived can be very costly for the agent as illustrated in Figure 3. Furthermore, this cost is borne by agents with a positive state who need to falsify on the equilibrium path. Therefore a planner who is even slightly concerned with the payoff of the agent might dislike this test. Finally, such cheating may have a detrimental effect on society by encouraging more cheating behavior, as documented in the works we mentioned in the introduction. For these reasons, it may be socially preferrable to employ only tests that do not generate any falsification. Motivated by this, we now focus on falsification-proof tests.

We characterize a receiver-optimal falsification-proof test for cost functions that satisfy either (UID) of (UDD). To keep the exposition simple, we focus on the case $\mu_\pi \leq 0$. We conclude the section by briefly explaining how to extend our analysis to the program of a planner with objective function $V(\tau, \phi) + \alpha U(\tau, \phi)$, and any value of μ_π. This more general characterization result is stated as Theorem B.1 in Online Appendix B. Throughout this section, we assume that $c(t|s)$ is regular in the sense of Definition 1, and that the prior is atomless.

Building on the recommendation principle and the equivalence of interim and ex ante falsification, we can write the receiver program as:[21]

$$\sup_{\tau} \int_S s\tau(s) dF_\pi(s) \tag{FPRP}$$

$$\text{s.t. } \tau(t) - \tau(s) \leq \gamma c(t|s), \quad \forall s, t \in S. \tag{FPIC}$$

We characterize a solution of this program under two distinct assumptions on the cost function. The first assumption, (UDD), ensures that we can use a first-order approach. The second one, (UID), allows us to connect this program with a well known optimal transportation problem. In both cases, we start by using a Lagrangian approach that circumscribes the problem to a single interval of states over which the test function is nondecreasing from 0 to some upper bound.[22]

[21] The receiver obedience constraint is still redundant as the uninformative test is falsification-proof.

[22] More accurately, the Lagrangian approach allows us to guess such an interval. We then rely

Restriction to smooth nondecreasing tests. We first show that we can restrict attention to Lipschitz and nondecreasing test functions (see Lemma B.3). Indeed, Lipschitz continuity is implied by (FPIC) and (REG). For monotonicity, we replace any falsification proof test by the highest monotonic function below it for negatives states, and the lowest monotonic function above it for nonnegative states. This generates a monotonic test that is preferred by the receiver, and preserves falsification proofness.

Therefore, we can work with tests that are almost everywhere differentiable with derivative τ' bounded in $[0, \gamma K]$, and such that, for every $s \in S$, $\tau(s) = \underline{\tau} + \int_{-\underline{s}}^{s} \tau'(z)dz$. So, instead of optimizing on the function τ, we can optimize on the scalar $\underline{\tau} \in [0, 1]$ and the function $\{\tau'(s)\}_{s \in S}$.

Differential Program. Then, we use integration by parts to rewrite the objective function in (FPRP) as

$$\underline{\tau}\mu_\pi + \int_S \tau'(z)J(z)dz,$$

where $J : S \to \mathbb{R}$ is defined by

$$J(z) = \int_z^{\overline{s}} sdF_\pi(s),$$

and is easily seen to be negative for $z < s_0$, nonnegative otherwise, continuous, increasing on S^-, decreasing on S^+, and therefore single-peaked at 0.

In addition, we face the *probability constraint* that τ must be bounded from above by 1, which we can rewrite as $\underline{\tau} + \int_S \tau'(z)dz \leq 1$, and the incentive constraint that, for every $s < t$, $\int_s^t \tau'(z)dz \leq \gamma c(t|s)$. Decreasing $\underline{\tau}$ weakly increases the objective function as $\mu_\pi \leq 0$, relaxes the probability constraint, and has no effect on the incentive constraints, implying that it is optimal to set $\underline{\tau} = 0$.

Lagrangian. Next, we treat the probability constraint with the Lagrangian method, with corresponding Lagrange multiplier $\lambda \geq 0$, which yields the Lagrangian objective:

$$\mathcal{L}(\tau', \lambda) = \int_S \tau'(z)J(z)dz + \lambda\left(1 - \int_S \tau'(z)dz\right) = \int_S \tau'(z)\big(J(z) - \lambda\big)dz + \lambda.$$

on a Lagrange sufficiency result established in Lemma 3 to show that our solution is optimal.

The Lagrangian problem is to maximize $\mathcal{L}(\tau', \lambda)$ where $\tau' : S \to [0, K]$ is feasible if, for every $s < t$, $\int_s^t \tau'(z)dz \leq \gamma c(t|s)$, and $\int_{s_0}^{\overline{s}} \tau'(z)dz \leq 1$. Clearly, any solution to this Lagrangian problem must satisfy $\tau'(s) = 0$ for almost every s such that $J(s) < \lambda$, that is, by continuity and single-peakedness of J, outside of an interval $[s_*(\lambda), s^*(\lambda)]$ such that $J\big(s_*(\lambda)\big) = J\big(s^*(\lambda)\big) = \lambda$, where $s_0 \leq s_*(\lambda) \leq 0 \leq s^*(\lambda)$ are uniquely defined. Furthermore, note that

$$\int_{s_*(\lambda)}^{s^*(\lambda)} s dF_\pi(s) = J\big(s^*(\lambda)\big) - J\big(s_*(\lambda)\big) = 0. \tag{1}$$

The following result combines these observations with a Lagrangian sufficiency result.

Lemma 3. *Suppose that there exists a Lagrange multiplier $\lambda \geq 0$, and a test $\tau : S \to [0, 1]$ that is nondecreasing and γK-Lipschitz such that:*

(i) $\tau(s) = 0$ *for every* $s \leq s_*(\lambda)$.

(ii) $\tau(s) = \tau\big(s^*(\lambda)\big)$ *for every* $s \geq s^*(\lambda)$

(iii) $\tau(t) - \tau(s) \leq \gamma c(t|s)$ *for every* $s_*(\lambda) \leq s < t \leq s^*(\lambda)$.

(iv) $\lambda = 0$ *or* $\tau\big(s^*(\lambda)\big) = 1$.

(v) For every nondecreasing and γK-Lipschitz test $\hat{\tau} : S \to [0, 1]$ that satisfies (i)-(iii),

$$\int_{s_*(\lambda)}^{s^*(\lambda)} s\tau(s)dF_\pi(s) \geq \int_{s_*(\lambda)}^{s^*(\lambda)} s\hat{\tau}(s)dF_\pi(s).$$

Then τ is a receiver-optimal falsification-proof test.

A Matching Function. For each $s_* \in [s_0, 0]$, there is a unique $s^* = m(s_*)$ in $[0, \overline{s}]$ such that $J(s_*) = J(s^*)$, where the decreasing *matching function* $m : [s_0, 0] \to [0, \overline{s}]$ is implicitly defined by $J(s) = J(m(s))$, or equivalently by $\int_s^{m(s)} z dF_\pi(z) = 0$. In particular, s_0 is matched to $m(s_0) = \overline{s}$. To each $s_* \in [s_0, 0]$ corresponds a single Lagrange multiplier $\lambda = J(s_*) \geq 0$, such that $s_*(\lambda) = s_*$. This matching function plays an important role in the characterization of the optimal test.

Next, we use two distinct approaches to derive an optimal test under the (UDD) and (UID) assumptions.

First-Order Approach. Following the Lagrangian method, we start by choosing a value of the Lagrange multiplier λ, and an optimal test τ for which we show that the conditions of Lemma 3 hold. The first-order approach consists in replacing the set of incentive constraints by first-order conditions. Relying on Lemma 3, we focus on nondecreasing test functions that satisfy (i) to (iii). Then, the first order condition for $\tau(t) - \gamma c(t|s)$ to be maximized at $t = s$ is

$$\tau'(t) \leq \gamma c_{t^+}(s|s), \tag{FPICFOC}$$

where $c_{t^+}(s|s)$ denotes the right derivative of $c(t|s)$ with respect to t at $t = s$. And to maximize the Lagrangian, which is the same as satisfying (v), we should intuitively set $\tau'(t) = \gamma c_{t^+}(s|s)$ for $s \in [s_*(\lambda), s^*(\lambda)]$ since that is exactly where $J(s) \geq \lambda$. This leads to the test

$$\tau(s) = \begin{cases} 0 & \text{if } s < s_*, \\ \gamma \int_{s_*}^{s} c_{t^+}(z|z)dz & \text{if } s \in [s_*, s^*], \\ \gamma \int_{s_*}^{s^*} c_{t^+}(z|z)dz & \text{if } s \geq s^*. \end{cases}$$

To satisfy the complementary slackness condition (iv), we must ensure that $\tau\big(s^*(\lambda)\big) = 1$ or $\lambda = 0$. Given the form of the candidate optimal test, the first condition is $\int_{s_*(\lambda)}^{s^*(\lambda)} \gamma c_{t^+}(s|s)ds = 1$. This suggests choosing the Lagrange multiplier $\lambda_{fo} = \min\left\{\lambda \geq 0 : \int_{s_*(\lambda)}^{s^*(\lambda)} \gamma c_{t^+}(z|z)dz \leq 1\right\}$, and interval ends $s_* = s_*(\lambda_{fo}) = \min\left\{s \in [s_0, 0] : \int_{s}^{m(s)} \gamma c_{t^+}(z|z)dz \leq 1\right\}$, and $s^* = m\big(s_*\big)$.

Finally, for the first-order approach to be valid, we must ensure that the optimal test derived under the relaxed program obtained by replacing the incentive constraints (FPIC) by (FPICFOC), does satisfy (FPIC). A sufficient condition for this is that, for all $t \geq s$,

$$\tau'(t) - \gamma c_t(t|s) = \gamma c_{t^+}(s|s) - \gamma c_t(t|s) \leq 0,$$

which is equivalent to (UDD) for cost functions that satisfy (REG).

Optimal Transport Approach. In this case, we assume that the cost function satisfies (UID), and we show that the optimal test can then be obtained by drawing a connection with the theory of optimal transport.

Assume first that λ is given, and let $s_* = s_*(\lambda)$ and $s^* = s^*(\lambda) = m(s_*)$.

Lemma 3 suggests that we focus on the program of maximizing $\int_{s_*}^{s^*} s\tau(s)dF_\pi(s)$ over tests that satisfy all the requirements of the lemma up to (iv). However, we start by solving a relaxed program, keeping as our only constraint on τ the no falsification constraint from *negative* states to *positive* states

$$\begin{aligned}&\max_{\tau} \int_{s_*}^{s^*} s\tau(s)dF_\pi(s)\\ &\text{s.t. } \tau(t)-\tau(s) \leq \gamma c(t|s), \quad \forall s_* \leq s \leq 0 \leq t \leq s^*.\end{aligned}$$

To draw the connection with optimal transport, we also change variables and let $y=-s \in Y=[0,-s_*]$ and $z=t \in Z=[0,s^*]$. Finally, we let $\rho: Y \to \mathbb{R}$, and $\psi: Z \to \mathbb{R}$ be the functions defined by $\rho(y)=\tau(-y)=\tau(s)$, and $\psi(z)=\tau(z)=\tau(t)$. With these notations, the remaining incentive constraints become

$$\psi(z)-\rho(y) \leq c(z|-y), \quad \forall (y,z) \in Y \times Z.$$

And, up to multiplication by the constant $\mu^* = \int_0^{s^*} sdF_\pi(s) = \int_{s_*}^0 sdF_\pi(s)$, the objective function of the program becomes

$$\int_Z \psi(z)dQ(z) - \int_Y \rho(y)dP(y),$$

where $Q(z)=\frac{1}{\mu^*}\int_0^z xdF_\pi(x)$, and $P(y)=\frac{1}{\mu^*}\int_0^y xdF_\pi(-x)$ define atomless cumulative distribution functions on, respectively, Z and Y.

To summarize, the new relaxed and reformulated program is

$$\begin{aligned}&\sup_{\rho,\psi} \int_Z \psi(z)dQ(z) - \int_Y \rho(y)dP(y)\\ &\text{s.t. } \psi(z)-\rho(y) \leq \gamma c(z|-y), \quad \forall (y,z) \in Y \times Z,\end{aligned}$$

which we recognize as the dual of the following well-known Monge-Kantorovich optimal transport problem

$$\inf_{\varphi \in \mathcal{M}(P,Q)} \gamma \int_{Z \times Y} c(z|-y)d\varphi(z,y),$$

where $\mathcal{M}(P,Q)$ is the set of joint distributions on $Z \times Y$ with marginals Q on Z, and P on Y.

To intuitively understand this connection, note that the primal optimal transport problem can be thought of as the problem of a central planner needing to transport at minimal cost a good produced in locations all over Y in amounts distributed according to P to locations all over Z in amounts distributed according to Q with a transportation cost between locations given by $\gamma c(z|-y)$. The dual problem can be thought of as the problem of a profit maximizing transporter with the technology to transport the good at no cost setting price $\rho(y)$ at which she buys the good in each location y, and price $\psi(z)$ at which she sells the good in each location z. To ensure that the central planner delegates each transportation operation to her, she needs to choose prices so that the planner's cost of delegating $\psi(z) - \rho(y)$ is less than the cost she would have incurred on her own $\gamma c(z|-y)$. In our problem, the test designer plays the role of the transporter, and passing probabilities play the role of prices that must be set so as to dissuade the agent (planner) to falsify (transport) negative states as (to) positive states.[23]

By (UID), the transportation cost function of this problem, $c(z|-y)$ is submodular, implying a well-known solution for both problems.[24] Rewriting this solution[25] in terms of our initial notations, and completing for states outside of $[s_*, s^*]$, we obtain the test

$$\tau(s) = \begin{cases} 0 & \text{if } s < s_*, \\ -\gamma \int_{s_*}^{s} c_s\big(m(x)|x\big)dx & \text{if } s \in [s_*, 0], \\ \gamma c(s^*|s_*) - \gamma \int_{s}^{s^*} c_t\big(x|m^{-1}(x)\big)dx & \text{if } s \in [0, s^*], \\ \gamma c(s^*|s_*) & \text{if } s \geq s^*. \end{cases}$$

Next, we need to guess the value of the Lagrange multiplier. In order to satisfy the complementary slackness condition (iv), we choose for λ the smallest possible value that makes $\tau\big(s^*(\lambda)\big) \leq 1$, that is, $\lambda_{ot} = \min\big\{\lambda \geq 0 : \gamma c\big(s^*(\lambda)|s_*(\lambda)\big) \leq 1\big\}$, leading to interval ends $s_* = s_*(\lambda) = \min\big\{s \in [s_0, 0] : \gamma c\big(m(s)|s\big) \leq 1\big\}$, and $s^* = m\big(s_*\big)$.

To ensure that this test solves the program derived from Lemma 3, we need to show that it satisfies the relaxed incentive constraints. In the proof of Theorem 2,

[23] This analogy suggests that the primal problem could also be related to the problem of an optimally falsifying agent in our context. We can indeed show that the problem of an agent optimally falsifying states under a fully revealing test in the observable case can be reformulated as a primal optimal transport problem.

[24] See, for example, Galichon (2018, Chapter 4).

[25] In fact, the solution to the dual Monge-Kantorovich problem is determined up to a constant which, for our purpose, we choose to ensure that $\tau(s_*) = 0$.

we show that this is ensured by (UID). Note that the solution of the primal optimal transport problem is given by the degenerate transport map that transports y to $Q^{-1}\big(P(y)\big) = m(-y)$. In terms of our original problem, this means that the only binding incentive constraints are those between source states $s \in [s_*, 0]$ and target states $t = m(s)$ in $[0, s^*]$ obtained by applying the matching function.

Optimal Test. The following theorem characterizes optimal falsification tests:

Theorem 2. *Suppose that π has full support and no atom on $S = [-\underline{s}, \overline{s}]$, with $\mu_\pi < 0$, and that c satisfies (REG). If c satisfies (UDD), the following is a receiver-optimal falsification-proof test:*

$$\tau_{fo}(s) = \begin{cases} 0 & \text{if } s < s_*, \\ \gamma \int_{s_*}^{s} c_{t^+}(z|z)dz & \text{if } s \in [s_*, s^*], \\ 1 & \text{if } s > s^*, \end{cases}$$

with $s_* = \min\Big\{s \in [s_0, 0] : \int_s^{m(s)} \gamma c_{t^+}(z|z)dz \leq 1\Big\}$*, and* $s^* = m(s_*)$*.*
If instead c satisfies (UID), the following is a receiver-optimal falsification-proof test:

$$\tau_{ot}(s) = \begin{cases} 0 & \text{if } s < s_* \\ -\gamma \int_{s_*}^{s} c_s\big(m(x)|x\big)dx & \text{if } s \in [s_*, 0], \\ \gamma c(s^*|s_*) - \gamma \int_s^{s^*} c_t\big(x|m^{-1}(x)\big)dx & \text{if } s \in [0, s^*], \\ 1 & \text{if } s > s^*, \end{cases}$$

where $s_* = \min\big\{s \in [s_0, 0] : \gamma c\big(m(s)|s\big) \leq 1\big\}$*, and* $s^* = m\big(s_*\big)$*.*

The optimal falsification-proof test is uninformative for any cost function that satisfies (UDD) with a null marginal cost of upward falsification, such as $c(t|s) = (t-s)^2$. A positive marginal cost, by contrast, can be leveraged so as to deliver some useful information to the receiver without inducing falsification. For example, the optimal test for $c(t|s) = a|t-s| + b(t-s)^2$, with $a > 0$, $b \geq 0$, is linear with slope γa on $[-s^*, s^*]$. Note that, for any cost function satisfying (UID), this marginal cost must be positive.

Generalization. Both methods can be used to derive tests that describe the whole FP-constrained Pareto frontier. Indeed, the program of a planner putting

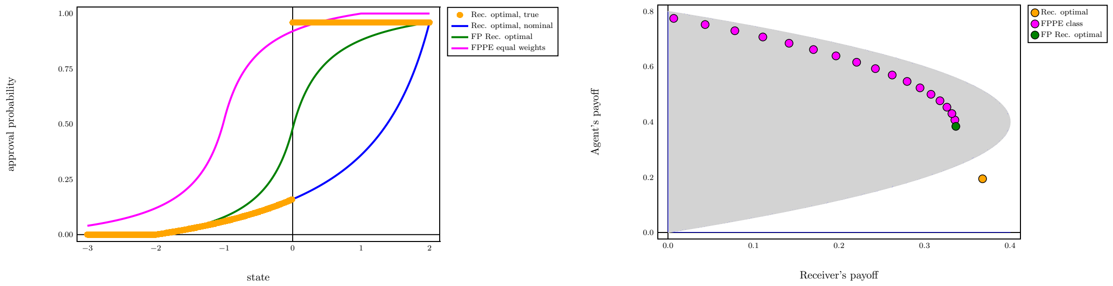

Figure 4: *On the left panel, we plot: the receiver-optimal test, and the true approval probabilities it generates; the receiver-optimal falsification-proof test, and the optimal falsification-proof test for a planner putting equal weights on the receiver and the agent. On the right panel, we plot the payoffs associated with each of these tests in the payoff space, as well as the constrained Pareto frontier of falsification-proof tests, obtained by varying the weight α on the agent. The grey area depicts the set of attainable payoffs under all possible information structures and no falsification. All plots are for $\gamma c(t|s) = 1.2|t-s|/(1+|t-s|)$ if $t \geq s$, and $\pi = Uniform(-3,2)$.*

weight 1 on the receiver and $\alpha \geq 0$ on the agent is given by

$$\sup_{\tau} \int_S (\alpha + s)\tau(s)dF_\pi(s) \tag{FPPP}$$

$$\text{s.t. } \tau(t) - \tau(s) \leq \gamma c(t|s), \quad \forall s,t \in S \tag{FPIC}$$

$$\int_S s\tau(s)dF_\pi(s) \geq \max\{\mu_\pi, 0\}. \tag{FPRO}$$

We need to add the receiver obedience constraint, as it is no longer automatically satisfied at the solution of the relaxed problem. But we can proceed by solving the relaxed problem as above, and then checking whether the obedience constraint is satisfied. Since the relaxed planner's program is essentially a receiver's program with an ideal approval threshold at $-\alpha$ instead of 0, we can solve it by following the same steps. The details, as well as the ensuing characterization results are stated in Theorem B.1 of Online Appendix B. The right panel of Figure 4 illustrates how this characterization result allows us to depict the Pareto frontier for falsification-proof tests by varying α.

Implications. For cost functions satisfying (UID), we can compare the receiver-optimal falsification-proof to the receiver-optimal test.

Proposition 3. *Suppose that c satisfies (CTT), (UID) and (REG), and that π has full support, and no atom on $S = [-\underline{s}, \overline{s}]$. Then the receiver is strictly better off, and the agent strictly worse off, under the receiver-optimal test than under the receiver-optimal falsification-proof test: $V(\tau_{ot}, \delta) < V(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*})$, and*

$U(\tau_{ot}, \delta) > U(\tau_{p^*,\hat{s}^*}, \phi_{p^*,\hat{s}^*})$. Furthermore, τ_{ot} and the corresponding payoffs of the receiver and the agent converge to the receiver's first-best as $\gamma \to \infty$.

Figure 4 depicts an example of such comparisons. While the receiver-optimal test is naturally better for the receiver, we prove that this ordering is strict, and we also show that the reverse strict ordering holds for the agent. The fact that the agent is better off under the falsification-proof test suggests implies that opting for such a test might help a planner worried about the possibly heavy cost of the receiver-optimal test to the agent. The right panel of Figure 4 shows how a planner worried by the externality of falsification, and with a positive weight on the agent could choose a test.

4 Observable Falsification

In this section, we study equilibrium information structures when falsification is observable directly or detectable. In this case, the receiver's posterior beliefs reflect actual rather than anticipated falsification. Deviations from the equilibrium falsification strategies may therefore lead the receiver to revise the conditional mean associated with a given signal downward (*devaluation*), or upward (appreciation), and revise her action as a consequence. This channel gives the test designer a new tool to deter deviations by ensuring that they lead to detrimental devaluations. However, this also makes the analysis of observable falsification more complex, and we therefore restrict our analysis to the binary state case $S = \{-\underline{s}, \overline{s}\}$, with $\mu_\pi = \pi\overline{s} - (1-\pi)\underline{s} < 0$.

We let $\underline{\phi} = \phi(\overline{s}|-\underline{s})$, $\overline{\phi} = \phi(-\underline{s}|\overline{s})$ and as before $\underline{c} = \gamma c(\overline{s}|-\underline{s})$, and $\overline{c} = \gamma c(-\underline{s}|\overline{s})$. We start by characterizing the receiver-optimal test under the assumption that the agent can only falsify upward, that is, by exogenously setting $\overline{\phi} = 0$, in Theorem 3. We then extend this result by providing a necessary and sufficient on the costs of upward and downward falsification for this test to remain optimal in Proposition 5.

We have illustrated in Example 1 that a fully informative test[26] is falsified in a way that makes the receiver indifferent between approving and rejecting, thus giving her a null payoff. Example 3 illustrates why, when ϕ is observable, enriching a test with an additional passing signal can make the receiver better off: The third signal enables the test to leverage the devaluation effect to prevent falsification while at the same time generating approval for the positive state with

[26] In fact, *any* two-signal test that induces approval for the positive state.

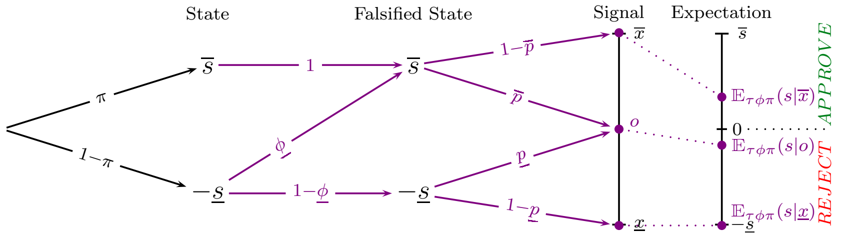

Figure 5: *This figure illustrates the three-signal test of Example 3. The expectation column shows how the mean associated with each signal shifts under falsification.*

probability one. In fact, adding a fourth signal or more increases the receiver's payoff even further.[27] The example also shows that the recommendation principle does not work with observable falsification, as it exhibits a three-signal test that yields a strictly positive payoff for the receiver, whereas any two-signal test yields zero to the receiver.

Example 3 (A three-signal test for the observable case). *Consider a test with discrete signal space $X = \{\underline{x}, o, \overline{x}\}$, 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:*

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

leading the receiver to approve after o and $\overline{x}$, and reject otherwise. Assume upward only falsification, so $\overline{\phi} = 0$. With $\underline{\phi} > 0$, we have:

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

Therefore, any positive falsification rate devalues signal o, leading the receiver to reject. The agent trades-off this negative effect of falsification with the positive effect of increasing the probability that signal $-\underline{s}$ generates signal $\overline{x}$. If the agent chooses $\underline{\phi} > 0$, he must ensure that $\mathbb{E}_{\tau\phi}(s|\overline{x}) \geq 0$ to induce the receiver to approve after signal $\overline{x}$, which yields $\underline{\phi} \leq \varphi_0$. The payoff of the agent for $0 \leq \underline{\phi} \leq \varphi_0$ is therefore given by

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

[27] Derivation details available upon request

Hence, setting $\overline{p} \geq \frac{\overline{s}(1-\underline{c})}{\underline{s}+2\overline{s}}$ ensures that the agent has no incentive to falsify. The receiver is then certain that the state is positive when she gets the high signal and is strictly better off under this test than with no information, or with an optimally falsified fully informative test. Furthermore, the receiver is better off with smaller values of $\underline{p}$ (and hence $\overline{p}$), as it lowers her probability of approving negative states. Therefore the best test she can pick in this class of falsification-proof tests 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 $\underline{c} = 0$. ◇

We proceed to derive a receiver-optimal test in closed form. Pushing the intuition of Example 3, this test uses a continuum of signals. Our characterization relies on the falsification-proofness principle (Proposition 1), and on the representation of tests as the distribution of conditional means they generate for the receiver, which amounts to relabelling signals as means.[28]

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.[29] We thus describe tests by the distribution of conditional expectations they generate, which amounts to normalizing signals as means. A test is therefore represented as a 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)

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

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. It is well-known that this representation is 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. As a consequence of this representation, we from now on equate signals with the posterior mean they generate given the test (and in the absence of falsification).

[28] Given the binary-state case, this amounts to saying that the belief-based approach is valid.

[29] See Lemma C.1 in Online Appendix C.

Rewriting Payoffs. The payoffs obtained by the receiver and the agent under $\mathcal{H}$, and in the absence of falsification, are respectively given by[30]

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

and

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

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. Next, we consider the effect of falsification on the receiver. Increasing $\underline{\phi}$ sends the negative state towards any signal $x \geq 0$ 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. Hence, 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 is independent of the test.

Lemma 4. *If $\underline{\phi} > \varphi_0$, all signals lead to rejection. If $\underline{\phi} \leq \varphi_0$, there exists a threshold $\hat{x}(\underline{\phi}) = \frac{-\mu_\pi \underline{s} \underline{\phi}}{\pi(\overline{s} + \underline{s}) - \underline{\phi} \underline{s}}$ such that the receiver approves for signals $x \geq \hat{x}(\underline{\phi})$, and rejects otherwise.*

This result implies that falsification levels outside of $[0, \varphi_0]$ are dominated for the agent. Furthermore, because there is a one-to-one relationship between any $\underline{\phi}$ in this range and the threshold it generates on $[0, \overline{s}]$, we can reformulate the receiver's falsification problem as the choice of an approval threshold[31] $x \in [0, \overline{s}]$ 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)**.** *An equilibrium is characterized by an approval threshold $x \in [0, \overline{s}]$ for the receiver, and a falsification level $\underline{\phi} \in [0, \varphi_0]$*

[30] We slightly abuse notations and use the same payoff functions as above with the new representations of tests and falsification. The second expression for the receiver's payoff is obtained using integration by parts.

[31] With a slight abuse of notation, we denote this threshold by x, as each nonnegative signal can be induced as a threshold by some falsification strategy.

such that $\underline{\phi} = \hat{\phi}(x)$, and x maximizes the agent's payoff

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

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 respectively generated by the negative and positive state to perform this computation. They are respectively given by the cdfs[32]

$$\overline{H}(x) = \frac{1}{\mu_\pi + \underline{s}}\big\{(x + \underline{s})H(x) - \mathcal{H}(x)\big\}, \tag{2}$$

and

$$\underline{H}(x) = \frac{1}{\overline{s} - \mu_\pi}\big\{(\overline{s} - x)H(x) + \mathcal{H}(x)\big\}. \tag{3}$$

The Receiver Program. Using the falsification-proofness principle, we can now reformulate the program for finding a receiver-optimal test as that of choosing a test function $\mathcal{H} \in \Delta^B$ to maximize $\mathcal{H}(0)$, under the falsification proofness constraint that 0 is an equilibrium threshold:

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

Using the expression of the agent's payoff in Proposition 4, the constraint can be rewritten as:

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

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

Linearization for Negative States. A first remark is that we can focus on test functions $\mathcal{H}$ that are linear on $[-\underline{s}, 0]$. Indeed, for any test function $\mathcal{H} \in \Delta^B$

[32] To see that, note that the joint probability that the state is positive 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 positive 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.

that satisfies (OFPIC), 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 since $\tilde{H}_\ell(0) = \mathcal{H}(0)/\underline{s} \leq H_\ell(0)$ by convexity of $\mathcal{H}$, satisfies (OFPIC) by the same argument, and is linear below 0. Going back to the interpretation of test functions, this implies that we can focus on tests that do not generate any signal (posterior mean) on $(-\underline{s}, 0)$. Such tests therefore have a single rejected signal, generated by the negative state only, with associated posterior mean $-\underline{s}$.

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. Denoting by κ its slope below 0, this test function must solve the indifference differential equation[33]

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

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

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

where

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

A Receiver Optimal Test. We show that $\mathcal{H}^*_{\underline{c}}$ is in fact receiver-optimal. 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 $-\underline{s}$,

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

which is only generated by the low type. $\mathcal{H}^*_{\underline{c}}$ puts an atom of size $\kappa^*_{\underline{c}}$ on this signal, and makes the agent indifferent across all the new 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. To see that, note that if $\mathcal{H}$ is a test which puts an atom of size $\kappa > \kappa^*_{\underline{c}}$ on the rejected signal, there must exist a signal x' between 0 and $\overline{s}$ such that $\mathcal{H}$ first crosses $\mathcal{H}^*_{\underline{c}}$ from above at x'. 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 implies that the agent prefers inducing falsification threshold x' to not falsifying under $\mathcal{H}$.

Theorem 3. *$\mathcal{H}^*_{\underline{c}}$ is the unique test function that solves (IDE) on $[0, \overline{s}]$, and it is a receiver-optimal test function under upward-only falsification. It is strictly increasing in $\underline{c}$ in the Blackwell informativeness order, and converges to the fully informative test function as $\underline{c} \to 1$. As a consequence, the payoff of the receiver is also strictly increasing in $\underline{c}$. Furthermore, $\mathcal{H}^*_{\underline{c}}$ is more Blackwell informative than any other receiver-optimal test function at $\underline{c}$. Finally, it is also Pareto efficient and delivers at least half of the receiver's payoff under full information, and this bound is tight when $\underline{c} = 0$.*

The efficiency of receiver optimal tests offers a stark contrast with the unobservable case, as illustrated on Figure 6. Making the falsification choices of the agent observable, or, equivalently, giving him the means to commit to his falsification strategy, leads to an efficient outcome even if falsification is costless. Furthermore, compared to the Bayesian persuasion benchmark, where the agent can commit to any information structure, or to the unobservable falsification case, which both lead to a null payoff for the receiver under costless falsification, our receiver optimal test restores at least half of her first-best payoff.

A general condition on costs. 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 5. *There exists constants $A > 0$ and B such that the test $\mathcal{H}^*_{\underline{c}}$ is receiver-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 $(\underline{\phi}, 0)$, as it leads the receiver to use a threshold $\hat{x} \geq \hat{x}(\underline{\phi})$, while lowering the probability that the positive state generates passing signals. Since $(\underline{\phi}, 0)$ is, by construction, not profitable, this is also the case of $(\underline{\phi}, \overline{\phi})$. 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 $\pi\overline{c} + (1-\pi)(1-\varphi_0)\underline{c}$. By comparing this payoff to the truth-telling payoff $1 - \kappa^*_{\underline{c}}$, we obtain the condition of the proposition.

Properties of Optimal Test. The following proposition derives the key properties of our optimal test. We depict its conditional and unconditional CDF and densities in Figure 6.

Proposition 6. *$H^*_{\underline{c}}$ 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 $\underline{H}^*_{\underline{c}}$.*

Our receiver-optimal test has a continuum of passing signals in spite of the binary-type and binary-action environment. In contrast, only one signal is associated with failure. There is a clustering of signals close to the threshold as illustrated on Figure 6. Furthermore, it makes the agent indifferent across all undominated levels of falsification[34] as it satisfies (IDE). The richness in passing signals, as well as the shape of the test, which is dictated by indifference, jointly contribute to maximizing the implicit (endogenous) falsification cost at every falsification level. Increases in $\underline{\phi}$ translate in devaluation of previously approved signals. When a signal is missing, the falsification level that would make this signal the new approval threshold is strictly dominated. By adding such a signal, the test can make the agent's incentive constraint bind at this falsification level, and increase the receiver's payoff. Roughly, the more passing signals there are, the more incentive constraints are made binding. A higher $\underline{\phi}$ increases the probability that the negative state generates the continuum of passing signals. But the receiver reacts by rejecting for some of the previously approved signals in an amount

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

that exactly offsets the advantage from the first effect. This signal devaluation is the main lever of the optimal test and works precisely because ϕ is observable.

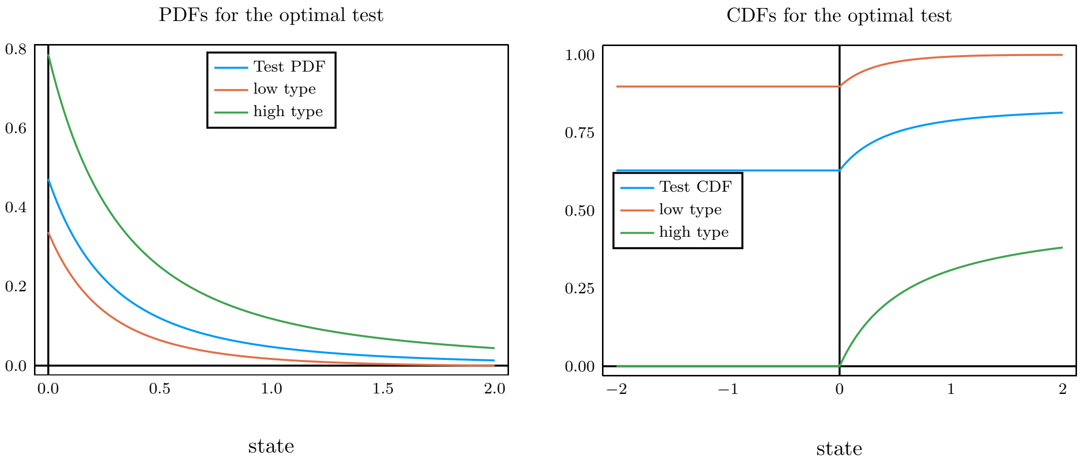

Figure 6: *PDF and CDF of the receiver-optimal test under observable falsification.* $-\underline{s} = -2$, $\overline{s} = 2$, $\pi = 0.3$.

Receiver commitment is valuable under observable falsification. In Proposition 2, we established that when falsification is unobservable commitment to the decision on the receiver's side is no better than commitment to a test. The opposite is true in the case of observable falsification. If the receiver can commit to reject regardless of the signal realization following any falsification, the only best response for the agent is to not falsify at all regardless of the test, and full information is equilibrium feasible.

Simple-versus fully optimal test: relative performance. As we establish in the proof of Theorem 3, both our optimal test and the three-signal test from Example 3[35] deliver at least 50% of the full information payoff to the receiver. A numerical analysis we perform in Perez-Richet and Skreta (2018) shows that the three-signal test delivers at least around 80% of the optimal receiver payoff, suggesting that most of the benefits can be harvested with simple tests using a small number of signals. Figure 7 depicts the payoff vectors resulting from receiver-optimal tests with three (3S) and four (4S) signals, as well as the payoff vector for the receiver-optimal test of Theorem 3. All these payoff vectors are on the

[35] We can show that this test is in fact the optimal three-signal test. See Perez-Richet and Skreta (2018).

Pareto frontier and the receiver's payoff increases as the number of passing signals increases.

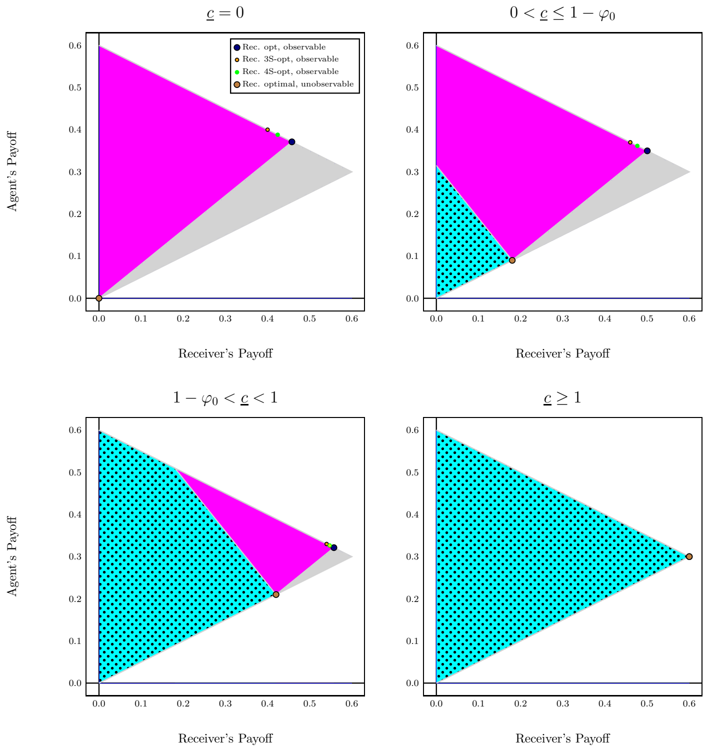

Figure 7: *The grey triangle depicts the space of feasible payoffs without falsification. The blue dotted area depicts the set of feasible payoffs under unobservable falsification. The pink area shows some of the additional payoffs that are feasible under observable falsification..*

Covert vs. Observable Falsification. We finish by comparing the equilibrium outcomes arising under unobservable and observable falsification. In Figure 7, we depict feasible payoffs under observable and covert falsification, in the binary state case.[36] The set of feasible payoffs under covert falsification (in blue) is also achievable under observable falsification, as it is easy to see that the agent has no incentive to falsify any of the tests at its extreme point under observable falsification. The KG test τ_{KG}, whose payoffs lie at the top vertex of the grey payoff triangle, is falsification-proof under observable falsification, and therefore feasible. Finally, our receiver-optimal test is also feasible. This implies that all payoffs

[36] Our results do not allow us to pursue this comparison beyond the binary state case. However, it is easy to see that both the receiver-optimal equilibrium information structure, and the receiver-optimal falsification-proof tests are feasible under observable upward only falsification.

in the pink area are also feasible under observable falsification. Furthermore, we know that no payoff vector to the right of the receiver-optimal test payoff vector is feasible. Overall, this shows that making falsification observable, or equivalently giving the agent the means to commit to his falsification strategy, enlarges the set of feasible payoffs, and makes it possible to attain efficiency even when upward falsification is costless.

References

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

Ajzenman, N. (2018): "The power of example: Corruption spurs corruption," *Working Paper*.

Alm, J., K. M. Bloomquist, and M. McKee (2017): "When you know your neighbour pays taxes: Information, peer effects and tax compliance," *Fiscal Studies*, 38, 587–613.

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*.

Baruchson-Arbib, S. and J. Bar-Ilan (2007): "Manipulating search engine algorithms: the case of Google," *Journal of Information, Communication and Ethics in Society*.

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.

Bloedel, A. W. and I. Segal (2020): "Persuading a Rationally Inattentive Agent," *Working Paper.*

Boleslavsky, R. and K. Kim (2018): "Bayesian persuasion and moral hazard," *Working Paper.*

Brynjolfsson, E. and T. Mitchell (2017): "What can machine learning do? Workforce implications," *Science*, 358, 1530–1534.

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.*

Du, S. (2018): "Robust mechanisms under common valuation," *Econometrica*, 86, 1569–1588.

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.

——— (2020): "Improving information from manipulable data," *Working Paper.*

Galbiati, R. and G. Zanella (2012): "The tax evasion social multiplier: Evidence from Italy," *Journal of Public Economics*, 96, 485–494.

Galichon, A. (2018): *Optimal transport methods in economics*, Princeton University Press.

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

Gneezy, U., A. Kajackaite, and J. Sobel (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–53.

Guo, Y. and E. Shmaya (2020): "Costly miscalibration," *TE*, forthcoming.

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.

Landier, A. and G. Plantin (2016): "Taxing the Rich," *The Review of Economic Studies*, 84, 1186–1209.

Lipnowski, E., L. Mathevet, and D. Wei (2020): "Attention Management," *Amercian Economic Review: Insights*, 2, 17–32.

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.

——— (1986): "Multistage Games with Communication," *Econometrica*, 54, 323–358.

——— (1991): *Game Theory, Analysis of Conflict*, Harvard University Press.

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.

Perez-Richet, E. and V. Skreta (2018): “Test design under falsification,” *Working Paper.*

Rincke, J. and C. Traxler (2011): “Enforcement spillovers,” *Review of Economics and Statistics*, 93, 1224–1234.

Rodina, D. (2016): “Information Design and Career Concerns,” *Working Paper.*

Rodina, D. and J. Farragut (2016): “Inducing Effort through Grades,” *Working Paper.*

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.

Stevenson, M. T. and J. L. Doleac (2019): “Algorithmic Risk Assessment in the Hands of Humans,” *Working Paper.*

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