Test Design under Falsification*

Eduardo Perez-Richet† Vasiliki Skreta ‡

July 20, 2020

Abstract

We characterize a receiver-optimal test when manipulations are possible in the form of type falsification. When falsification unobservable, but costless, trivially, no test can generate information. When falsification is costly, the optimal test consists of a highest possible passing probability, p, and the smallest positive state that passes with that probability, s^*. All lower positive states below falsify to s^*, whereas all negative pass with a probability that makes them indifferent between falsifying to s^* and obtain the highest possibility approval minus the falsification costs, and not falsifying. There is a range of negative states that are not approved. Falsification and cheating can create negative externalities, so we characterize the optimal test that prevents falsification and it is worse: it approves negative states with higher probability while approves positive ones with lower probability. We also derive an optimal test when falsification is observable (or committed to) in a binary state setting with linear costs. The optimal test is falsification proof and "rich" in that it has a continuum of passing signals. It leverages the endogenous depreciation of signals (caused by observability) to generate information even if explicit falsification costs are zero. Both the agent and the receiver strictly benefit from falsification being observable or detectable. Keywords: Information Design, Falsification, Tests, Manipulation, Cheating, Persuasion.
JEL classification: C72; D82.

*We thank the Editor and anonymous referees for excellent comments. Ricardo Alonso, Philippe Jehiel, Ines Moreno de Barreda, Meg Meyer, Philip Strack and Peter Sorensen 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."

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

1 Introduction

Tests are prevalent, and stakes are often high for all concerned parties. Teachers prepare their students to pass tests in order to gain admission to selective schools and universities. Issuers seek to obtain a good rating for their assets. Pharmaceutical companies seek FDA's approval for new drugs. Car manufacturers need to have their vehicles pass emission tests. The list is suggestive of how wide-ranging and relevant tests are, and why it is important that test results are reliable: Fairness, inadequacy, financial distraught, and environmental pollution are at stake when tests are compromised.

However, manipulations are equally prevalent, and often successful. They are common in standardised graduate admission tests. Pharmaceuticals have come under scrutiny for using sub-standard clinical trial designs in order to obtain FDA's approval as in Sarepta's case (*The Economist*, October 15, 2016).[1] Car manufacturers sometimes cheat on pollution emission tests. Some manipulations can be socially acceptable and observable such as universities hiring part-time prominent scholars to increase their ranking,[2] or parents excessively tutoring their children. This is the first paper to study the optimal design of tests in the presence of manipulations.

We consider a agent-receiver relationship, in which the agent would like to convince the receiver to approve his items. The receiver—or several identical receivers, employers, investors, consumers each facing one item—wishes to approve items selectively, depending on their hidden state. In particular, no approval yields zero to the receiver, while approving an item of type s yields s. The prior mean of the state is negative, so without any additiobal information the receiver rejects. To uncover the types of the items, the receiver benefits from information generated by a test to which each item is subjected. This test is modeled as a Blackwell experiment: a probability distribution over signals (test results, grades) as a function of the type of an item. The receiver decides whether or not to approve after observing these signals, but cannot commit in advance to an approval policy contingent on signals.

The agent has a manipulation technology at his disposal. He can, possibly at a cost (explicit or psychological), falsify the type of some of his items for testing purposes, so that, for example, an item of quality s generates the same signal distribution as state t. A manipulation strategy is therefore a choice of falsification rates: with what probability an item of quality s is disguised

[1] http://www.economist.com/news/leaders/21708726-approving-unproven-drug-sets-worrying-precedent-bad-

[2] https://liorpachter.wordpress.com/2014/10/31/to-some-a-citation-is-worth-3-per-year/

any other quality t. Good illustrations of this manipulation technology are a teacher teaching a student to the test, or the way Volkswagen compromised emission tests.[3]

Typically falsification is unobserved and deviations are not detectable. We call this the *unobservable* case (and sometimes "non-committed" to contrast it with the case that follows). When the agent's deviations are undetected the receiver's belief after each signal is based on the "equilibrium" anticipated falsification rather than the actual. In contrast, when falsification is observable the "meaning" of signals reacts to deviations so a passing signal may become a failing one. Depending on the situation, both benchmarks are relevant so we derive receiver-optimal tests for both the observable and unobservable falsification settings and compare their features and payoffs they yield to the agent and the receiver. It is interesting to note that the insights we can leverage to establish the optimum in each of the two cases differ quite substantially.

When falsification unobservable, but costless, trivially, no test can generate additional information: if there is a test that passes some state at a higher rate than others, the agent will falsify all items as that state. When falsification is costly, the optimal test consists of a highest possible passing probability, p, and the smallest positive state that passes with that probability, s^*. These two parameters, together with falsification costs, dictate the smallest state that passes with positive probability. All positive states below s^* falsify to s^*, whereas all negative pass with a probability that makes them indifferent between falsifying to s^* and obtaining the highest possibility approval minus the falsification costs, and not falsifying. There is a range of negative states that are not approved.

We also characterize the optimal test that prevents falsification.[4] The shape of the optimal test is reflects the falsification costs and its derivation relies on identifying the optimal falsification target for each negative state. The target is a positive state, so non-local incentive constraints bind. The falsification proofness constraint binds for all negative states above a threshold, so there is a continuum of binding constraints. Mathematically, the program characterizing the optimal test in this case, is equivalent to a mechanism design problem, and in particular, an optimal allocation problem, without transfers and costly reporting. The lack

[3] On January 11, 2017, "VW agreed to pay a criminal fine of $4.3bn for selling around 500,000 cars fitted with so-called "defeat devices" that are designed to reduce emissions of nitrogen oxide (NOx) under test conditions." https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal

[4] The presence of falsification and more general other forms of cheating to documented to negative externalities to society see, for example, Galbiati and Zanella (2012); Ajzenman (2018); Alm, Bloomquist, and McKee (2017); Rincke and Traxler (2011).

of transfers, the fact that we cannot rely on the usual approach that focuses on the "relaxed problem" (which ignores the incentive constraint), as well as the fact that local IC does not imply global IC, renders previous solution methodologies unsuitable.[5] To solve the problem we relied on the following steps: TO ADD

When falsification is observable, information can be generated even if it is costless. Optimal design exploits the following manipulator trade-off: while falsification may lead to better grades, it devalues their meaning. We consider a version of the model where state is binary, one is negative $-\underline{s}$ and the other is positive $\overline{s} > 0$. We take the falsification cost function to be linear in the probability ϕ that $-\underline{s}$ falsifies as $\overline{s}$. For this setting, we show that optimal tests can be derived among falsification-proof ones. The optimal test has a single 'failing' grade, and a continuum of 'passing' grades. It makes the manipulator indifferent across all moderate levels of falsification. The positive state never fails, but the negative sometimes may pass. An optimal test delivers at least half of the full-information value to the receiver, even if the explicit cost of falsification is zero. A three-grade optimal test also performs well.

We now discuss some key features of our model. First, we look at the falsification technology. Falsification can only make the receiver less informed, in a Blackwell sense, but does not make every garble of the test attainable. For example, the falsification technology allows the agent to render *any* test uninformative. If the prior mean is positive so that the receiver approves when her belief is equal to the prior, making the test uninformative is actually the optimal choice of the agent. This is why, in what follows, we focus on the interesting case where the prior mean is negative. For a given test, however, the agent cannot generate all the information structures that are less Blackwell informative than this test. This limitation is what makes the test design problem interesting. Indeed, if the agent could generate any such garbling, then the optimal design problem would always result in the optimal information structure of the agent. The reason we picked this technology is because it is natural and fits well a number of examples mentioned. However, other manipulation choices interesting and relevant for other settings. We discuss this further in the related literature section.

Unobservable and non-committed falsification can be viewed as (costly) communication: The agent's faslification strategy is like the sender's reporting rule in a sender-receiver game

[5] Severinov and Tam (2019) solve a related problem, with costly reports but they have transfers and following a different approach leveraging Hamiltonians (see also Rochet (1985)).

where the sender is faced with an exogenous communication device, the test τ. As in the communication games, the receiver, observes the output message x, forms beliefs about the state given the test and sender's strategy and chooses an action ex-post. In readily follows from the revelation principle that staring from an abstract "test" $\tau :\to \Delta(X)$ with an arbitrary set of output messages X, one can take to be action recommendations. In our binary-action settings this implies that $X = \{$ "approve", "reject"$\}$. The proof follows from standard arguments making sure that in addition to inducing the receiver to choose the same action as in the original test, the agent does not have new falsification opportunities.

Committed or observable falsification can be viewed as constrained "persuasion." Indeed, here the agent is "constrained" by the test put in place by the receiver (or, for that matter, by another party who shares the same preferences as the receiver(s)), and chooses a *faslification/persuasion rule*. The set of feasible information structures are constrained by the test in place. The illustrative example in Section 2 highlights that action recommendations are insufficient when falsification is committed. In fact, we establish in Section 5 that even with two actions and two states the optimal test involves a *continuum* of different signals associated with "approve". Commitment (observable) falsification is better for everyone, since as we establish in Section 2 in the binary state example, even without costs we can design tests that approve all good items and some of bad ones. This is impossible with non-committed falsification. Hence, the agent benefits in the same way the sender benefits from commitment in the usual Bayesian persuasion case: Commitment to a falsification rule is analogous to the corresponding assumption in (Kamenica and Gentzkow, 2011) where the agent chooses and commits to a test.

Commitment can be justified in a number of ways. Falsification rates can be inferred from the empirical distribution of grades if falsification strategy is chosen once and for all and used for multiple items. We explore the limit version of this argument by looking at the case of a continuum of items in Perez-Richet and Skreta (2018). It is also possible that the chosen falsification strategy is applied to multiple items that are tested sequentially allowing test users to learn the falsification strategy, either because the type of each item is revealed at the end of a period, or by looking at the distribution of past grades. In the case of a single item, falsification is a probability. This does not preclude observation as this probability may be the consequence of observable actions such as an effort or an investment. Also, even in the case of socially unacceptable manipulations, information about the level of manipulations may leak

and become publicly known because of bragging, whistleblowing or mere conversations.

Finally, we comment on the lack of commitment assumption by receivers. When falsification is unobservable, the receiver cannot benefit from commitment it follows simply from our analysis that the receiver does not benefit from commitment. With observability, it would be possible to generate perfect information by committing to reject items regardless of signals whenever manipulations are observed. Such commitment is often problematic in practice: In reality, employers, consumers, investors see test scores first, and only then decide which workers to hire, which assets to buy and so on. If receivers are aware of a limited amount of manipulation that is insufficient to lower their belief below approval threshold, they are unlikely to reject. Our framework can accommodate commitment by a regulator to punish manipulations. Such punishments are a particular case of falsification costs. Suppose, for example, that the regulator is willing to punish the agent when she observes manipulations, but that she would not go so far as to force any item to be rejected regardless of the signal generated, or that, in order to do so, she would have to provide justifications, whether legal or internal. Then the expected punishment would incorporate the probability that such justifications are available and can be written as a falsification cost. Unsurprisingly, if such costs are sufficiently high even the fully informative test is not manipulated.

OLD parts of intro While this manipulation technology allows the agent to garble the information generated by the test, and to turn any test completely uninformative, it does not make all garbles available.[6] This limitation of available garbles helps receivers only if the set of signals generated by the test is sufficiently rich. Indeed, we show that the agent can garble any sufficiently informative binary test (such as the fully informative one) into his optimal information structure. Hence, receiver-optimal tests must use more than two signals.

The model, while stylized, captures a key trade-off: manipulations can increase the rate of approval, by increasing the chance that "bad" items generate good test results, but, in excess, they can make test results so unreliable that they nullify approvals. So, even if manipulations bear no cost, or punishment, excessive manipulations can hurt the agent. A rational agent, therefore, manipulates moderately. Manipulability complicates test design, as one has to take into account how manipulations alter the information structure generated by the original test.

[6]If all garbles were attainable, the agent could garble any sufficiently informative test into his optimal information structure—the one he would pick if he were the information designer, thus making the test worthless.

Our analysis shows how receiver-optimal design can exploit the aforementioned trade-off to obtain informative tests in spite of manipulations, even in the absence of explicit punishments or unrealistic commitment on the side of the receiver.[7]

The receiver-optimal test we derive has a number of remarkable features and delivers some practical insights. First, it is manipulation-proof in the sense that all agent types find it optimal to choose falsification rates equal to zero. Second, despite the fact that there are only two actions to take, it is "rich" in the sense that it generates a continuum of signals that lead to approval and only one that leads to rejection. Hence, the *receiver side* revelation principle that usually holds in Bayesian persuasion (Kamenica and Gentzkow, 2011) and mediation problems (Myerson, 1991, Chapter 6), which allows to reduce the information design problem to the problem of designing a recommendation system, does not hold in our environment. Third, all items that would be approved under full information are approved under the receiver-optimal test, but some items that should be rejected are also approved. That is, the optimal test leads to some false positives, but no false negatives. Fourth, it is ex-ante Pareto efficient, and gives the receiver at least 50% of the payoff she would get under full information. Fifth, the distribution of signals generated by the good type first-order stochastically dominates that generated by the bad type. Furthermore, our optimal test makes the agent indifferent between not manipulating, and any other approval threshold he could induce through manipulations.

To see why tests with more signals can be beneficial, it is useful to consider adding a third "noisy" signal to the fully informative test. We can choose the probabilities that the good and bad type generate this signal so that, in the absence of manipulations, it leads the receiver to a belief equal to the approval threshold $\hat{\mu}$. With such a test, any amount of falsification leads the receiver to lower the belief associated with the intermediate signal, and thus reject items that generate this signal. Then the agent has to weigh the benefit of manipulating (bad types are more likely to generate the top signal), with its endogenous cost (losing the mass of good and bad types that generate the intermediate signals). To make such a test as good as possible for the receiver, we can pick the test so that these two effects compensate each other, thus making the agent indifferent between his optimal amount of falsification, and no falsification. The resulting test is manipulation-proof, and generates valuable information for the receiver.

[7]With commitment or with richer contracts (or mediation schemes) it is possible to achieve the receiver-first-best in our model. We focus on test-design given the prevalence of tests, and given that they perform very well even without commitment on the side of the receiver.

In fact, we establish a general *no-falsification principle*, which shows that, for any test, there is an equivalent manipulation-proof test that generates the same information and payoffs to all parties. This result is a version of the revelation principle adapted to our environment. Combined with the representation of experiments as convex functions introduced in Kolotilin (2016), and further studied in Gentzkow and Kamenica (2016b), it allows us to reformulate the receiver-optimal design problem as a maximization problem over convex functions representing tests, under a no-manipulation incentive constraint. The no-manipulation incentive constraint can be formulated as a condition bearing on the payoff of approval thresholds induced by manipulations.

The optimal test we derive has a single signal associated with rejection generated by bad items only and it makes the agent indifferent between not manipulating, and inducing any other approval threshold through cheating. This test is characterized by a differential equation that we solve in closed form. We derive receiver-optimal tests under two conditions that we later relax: The first one is that falsification is perfectly observable, and the second is that falsification rates are constrained so that $p_B + p_G \leq 1$. The latter constraint rules out falsification rates so high that they would lead to an inversion of the meaning of signals. Both assumptions are useful in allowing us to focus on the main trade-offs, and are compelling in some cases but not always, so we show how to relax them in Section **??**

When manipulations are costly—the agent incurs a psychological or technological cost when manipulating, or is subject to fines when caught—the no-falsification principle holds if the marginal cost of increasing p_B does not increase too fast. We show that the fully informative test is optimal whenever the cost is sufficiently high. When it is not, we derive the optimal test under a linear cost function, and show that it satisfies the same properties as without cost. Furthermore, the receiver-optimal test becomes more informative as manipulations become more costly. In Appendix **??**, we show how to find an optimal test for a larger class of cost functions.

1.1 Related Literature

Manipulations in information design. Nguyen and Tan (2020) consider a Bayesian persuation setting where the sender privately observes the experiment's result and can send costly messages to the receiver. The cost function is a metric and it satisfies properties analogous

to the triangular inequality. In that paper the agent manipulates the experiment's *output*– whereas in our paper the experiment's input. There the designer, leverages the endogenous meaning of messages to assign expensive messages to desirable states. They show that sender-preferred equilibrium exists, then there exists an equilibrium where the Sender fully reveals. Li (2020a) is similar to Nguyen and Tan (2019), albeit in setting that is closer to that of Crawford and Sobel (1982); again the focus is on inducing full revelation by the agent.

Manipulations:

Hu, Immorlica, and Vaughan (2019) Zhang, Cheng, and Conitzer (2019a), Zhang, Cheng, and Conitzer (2019b) Guo and Shmaya (2018); Ball and Kattwinkel (2019) Frankel and Kartik (2019b,a)

Mechanism design with reporting costs. The papers Kephart and Conitzer (2016), Nguyen and Tan (2020) and Deneckere and Severinov (2017), investigate what is the shape of lying costs to get the revelation principle.

Deneckere and Severinov (2017) consider both costly signaling games and screening when signalling can happen with various dimensions (tests). Their main theorem shows that under certain assumptions (Assumption 1) on the costs structure, implementation can be achieved with almost no reporting costs–so there is essentially truth-telling in the limit. When they derive optimal mechanisms they impose sub modularity and increasing differences (assumption 2). aside: this paper has a very nice literature review Severinov and Tam (2019) consider a standard principal-agent screening model with transfers and fixed lying costs. In that setting given the nature of the costs and they fact that they have transfers, imply that they can restrict attention to direct mechanisms where truth-telling is a best response. They show that local IC is typically not binding–thus the standard approach of leveraging binding local IC to "replace" transfers and obtain a virtual surplus representation that only depends on the allocation seems inapplicable. The main contribution of the paper is to develop a method relying on optimal control theory to solve for the optimum. In our setting, we have no transfers, so the optimum involves falsification and our derivation of the optimal falsification-proof test cannot leverage the fact that there are transfers.

Kephart and Conitzer (2016) derive conditions that reporting costs must satisfy in order for get truth-telling without loss. They consider both transferable and non-transferable utility settings. A key condition is the triangular inequality that suffices with transferable utility.

Without transfers they require a stronger condition–for any ordered triplet a,b,c it must be the case that $ac \leq ab$ (FTVU), where they use $ab \equiv c(b|a)$ –the cost a incurs when reporting b. They have a counterexample to truthful implementation when this condition fails.

aside: this paper motivates the problem also with the rise of AI and algorithmic detection that uses big data to make inferences thus making lies harder / easier to detect.

Foundations of falsification and lying costs. There is ample empirical evidence that lying is costly, for a thorough study see Abeler, Nosenzo, and Raymond (2019). There are also works on identify the "shape" of costs and in particular documenting strictly positive marginal costs of lying Gneezy, Kajackaite, and Sobel (2018). The assumptions we make on the falsification cost functions are consistent with what is identified in that literature. Sobel (2020) explores lying in game theory.

Lying costs: Kartik (2009), Kartik, Ottaviani, and Squintani (2007),

Optimal allocation without transfers. There are several paper studying optimal assignment problems in the absence of transfers. They can be categorized according to the "tool" the designer leverages to elicit the agents' private information. One such tool is ex-post (costly) inspection or verification. This is the case in Ben-Porath, Dekel, and Lipman (2014); Lipman (2015); Mylovanov and Zapechelnyuk (2017); Chua, Hu, and Liu (2019); Epitropou and Vohra (2019); Li (2020b). Patel and Urgun (2017) add money burning (as in Condorelli (2012)) to the framework of Ben-Porath et al. (2014) with ex-ante identical agents. Both money burning and verification are part of the (Bayesian incentive compatible) optimal mechanism. Bhaskar and Sadler (2019) study the extent to which the designer, who wishes to maximize social surplus, can exploit the agents' limited preference alignment to extract information from the players. Guo and Hörner (2018) have an infinite discrete time horizon over which a principal and an agent interact. The authors solve for the optimal allocation mechanism. Unlike with transfers, efficiency decreases over time. Kattwinkel (2019) relies on the correlation to elicit information whenever possible and shows that actually in many cases random assignment is optimal. This is similar in spirit to Chakravarty and Kaplan (2013) where the agent can send a costly signal about type. The authors identify conditions under which ignoring these costly signals and using lotteries is optimal. In contrast to all these papers, in our paper there are reporting costs and we cannot rely on a revelation principle.

Theoretical work on Bayesian Persuasion. We introduce falsification in the information design literature. Kamenica and Gentzkow (2011) examine a party (sender) who wishes to design the best way to disclose information so as to persuade a decision-maker who may have different objectives.[8] In our paper the receiver chooses the experiment and the sender may tamper with the chosen experiment by falsifying the state.

We relate to recent works that study Bayesian persuasion in the presence of moral hazard. In Boleslavsky and Kim (2017), Rodina (2016), and Rodina and Farragut (2016), the prior distribution of the state is endogenous and depends of the agent's effort. The aforementioned papers differ in the principal's objective. Related to these works is Hörner and Lambert (2016), who find the rating system that maximizes the agent's effort in a dynamic model where the agent seeks to be promoted. In Rosar (2017) the principal designs a test that the agent decides whether or not to take. In our paper, participation to the test is not optional, and the agent cannot alter the distribution of types, but he can tamper with the test itself.

We also relate to Bizzotto, Rudiger, and Vigier (2016) and to Cohn, Rajan, and Strobl (2016), since there, like in our paper, certifiers designing tests need to take into account the fact that firms are not passive, but react to the certification environment. In Bizzotto et al. (2016) agents choose what additional information to disclose, whereas we investigate what happens when firms manipulate the information structure.

Our analysis is somewhat reminiscent to that of recent papers that study optimal information design in specific contexts. Chassang and Ortner (2016) design the optimal wage scheme to eliminate collusion between an agent and the monitor. The optimal wage scheme is similar to the buyer-optimal signal in Condorelli and Szentes (2016). In that paper as well as in Roesler and Szentes (2017), the buyer-optimal signal is such that the seller is indifferent across all prices he can set. Our paper uncovers a similar property, as the optimal test under observable falsification makes the agent indifferent across all moderate falsification levels.

On the technical side, we represent experiments as convex functions as in Kolotilin (2016) and Gentzkow and Kamenica (2016b). The latter study costly persuasion in a setup where the decision-maker cares only about the expectation of the state of the world. In our setup the receiver's decision also depends on a single-dimensional object: his belief that the state is good.

[8] There are several extensions of this leading paradigm including Gentzkow and Kamenica (2014), who allow for costly signals and Gentzkow and Kamenica (2016a) where two senders "compete" to persuade.

PUT THIS IN OBSERVABLE

Costly state falsification/Hidden income/Hidden Trades. Lacker and Weinberg (1989) incorporate costly state falsification in a risk-sharing model. Cunningham and Moreno de Barreda (2015) model manipulations as costly state falsification in a context similar to ours, but they study equilibrium properties under a fixed testing technology, whereas we focus on receiver-optimal test design. Hidden trades can also be viewed as a form of manipulation and are studied in Golosov and Tsyvinski (2007), and references therein. Grochulski (2007) models tax avoidance using a general income concealment technology analogous to the costly state falsification technology of Lacker and Weinberg (1989). In Landier and Plantin (2016), agents can hide part of their income which can be interpreted both as tax evasion and as tax avoidance.

2 Binary State Example

Throughout the paper, we consider a decision maker or receiver who can choose between two actions which we label *approve* and *reject* for simplicity. The rejection payoff to the receiver is normalized to 0, whereas the approval payoff is given by the state of the world. In this example, we consider a binary state of the world $s \in S = \{-\underline{s}, \overline{s}\}$, with $-\underline{s} < 0 < \overline{s}$. The prior probability of the high state is given by π_0, with $0 < \pi_0 < \frac{\underline{s}}{\overline{s}+\underline{s}}$, so that rejection is receiver-optimal at the prior. To take her decision, the receiver can rely on a test, that is a Blackwell experiment $\tau : S \to \Delta X$, where X is a measurable space of signals. We take this test as exogenously given and known to the players, and study how the test can make the receiver better or worse off. Normally, a fully informative test would be optimal for the receiver. But we assume that there is an agent who can manipulate the test by falsifying the state of the world that is fed to it. Specifically, in this example, the agent can choose the probability ϕ that state $\underline{s}$ generates signals according to $\tau(\overline{s})$ instead of $\tau(-\underline{s})$. In this example, falsification cost is linear, so the ex-ante cost of falsifying state $-\underline{s}$ as $\overline{s}$, with probability ϕ, is $(1-\pi_0)\phi c$. We assume that $0 \leq c < 1$, so that falsification may be worthwhile.

The agent chooses ϕ before knowing the state of the world. We discuss interim falsification choices in Section 3.

The fully informative test. Suppose first that the test is fully informative, so $\tau(-\underline{s})$ and $\tau(\overline{s})$ have disjoint support, which would be a receiver-optimal test in the absence of manipulation.

If ϕ is observable by the receiver she does not need to form beliefs about the choice of the agent, but takes ϕ into account when interpreting the results of the test. In particular, when seeing a signal that the state is high, $x \in \operatorname{supp} \tau(\overline{s})$, the receiver expects a payoff

$$\frac{\pi_0 \overline{s} - (1-\pi_0)\phi \underline{s}}{\pi_0 + (1-\pi_0)\phi}$$

from approving, which she therefore chooses if $\phi \leq \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}$. When seeing a signal $x \in \operatorname{supp} \tau(-\underline{s})$, the receiver is certain that the state is $-\underline{s}$ and rejects. The payoff of the agent is therefore given by

$$\pi_0 + \phi(1-\pi_0)(1-c)\, \mathbb{1}\left(\phi \leq \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}\right),$$

so the agent optimally chooses $\phi = \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}$, which is the falsification level that makes the receiver indifferent between both actions when receiving a signal indicative of the high state. The resulting information structure is agent-optimal and receiver-pessimal: It is the test the agent would optimally design if given the possibility (as in Kamenica and Gentzkow (2011)), and it gives the receiver the same null payoff as she would get without any information.

If instead falsification is not observable, the receiver must first form a belief about ϕ. However, all signals are on the equilibrium path for every choice of ϕ, so this belief is unaffected by the realized signal. Furthermore, it is correct in equilibrium. Because a signal in $\operatorname{supp} \tau(-\underline{s})$ can only be generated by the low state regardless of ϕ, such a signal always leads her 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 $\phi \leq \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}$. However, if the equilibrium strategy of the receiver is such that she approves for some signals, then the unique best response of the agent is to choose $\phi = 1$. Therefore the equilibrium strategy of the receiver must be to always reject. If $c = 0$, choosing any $\phi > \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}$ is a possible best response of the agent, and then this is an equilibrium in which both the agent and the receiver get their worst possible payoff. If $c > 0$, the only best response of the agent to the strategy of the receiver of always picking action 0 is to choose $\phi = 0$, but these are not mutual best-responses so an equilibrium does not exist.

A three-signal test for the observable case. Consider a test with discrete signal space $X = \{\underline{x}, o, \overline{x}\}$, and $\tau(\overline{s})$ is the probability distribution $(0, \overline{\tau}, 1-\overline{\tau})$, and $\tau(-\underline{s}) = (1-\underline{\tau}, \underline{\tau}, 0)$, with $\underline{\tau} = \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}}\overline{\tau}$. These values ensure 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 choose action 1 after signals o and $\overline{x}$, and 0 otherwise.

With falsification, we have:

$$\begin{aligned}
&\mathbb{E}_{\tau\phi\pi}(s|\overline{x}) \propto (1-\overline{\tau})\big(\pi_0\overline{s} - (1-\pi_0)\phi\underline{s}\big),\\
&\mathbb{E}_{\tau\phi\pi}(s|o) \propto \phi\big(\pi_0\overline{s} - (1-\pi_0)\underline{s}\big) \leq 0,\\
&\mathbb{E}_{\tau\phi\pi}(s|\underline{x}) = -\underline{s}.
\end{aligned}$$

Therefore, any positive level of falsification leads the receiver to reject following signal o. 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 $\phi > 0$, he must ensure that $\mathbb{E}_{\tau\phi}(s|\overline{x}) \geq 0$ to induce the receiver to choose approve after signal $\overline{x}$, which yields $\phi \leq \overline{\phi} = \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}$. The gain from falsification for the agent for $0 < \phi \leq \overline{\phi}$ is therefore given by

$$(1-\pi_0)\phi\big\{\mathbb{1}\left(\phi \leq \overline{\phi}\right)(1-\overline{\tau}) - c\big\} - \pi_0\overline{\tau} - (1-\pi_0)\underline{\tau}.$$

Hence, setting $\overline{\tau} = \frac{\overline{s}(1-c)}{\underline{s}+2\overline{s}}$, or $\overline{\tau} \geq 1-c$ ensures that the agent has no incentive to falsify. Because, the agent is then certain that the state is high when she gets the high signal, the receiver 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 $\underline{\tau}$ (and hence $\overline{\tau}$) so as to minimize her probability of picking action 1 in the low state. Therefore the best test she can pick in this class of falsification-proof tests is obtained by setting $\overline{\tau} = \frac{\overline{s}(1-c)}{\underline{s}+2\overline{s}}$.

A two-signal test for the unobservable case. Consider a test with binary signal space $X = \{\underline{x}, \overline{x}\}$, and conditional distributions $\tau(\overline{s}) = (1-c, c)$ and $\tau(-\underline{s}) = (1, 0)$. Then, it is an equilibrium for the agent to choose $\phi = 0$, and for the receiver to choose approve after signal $\overline{x}$

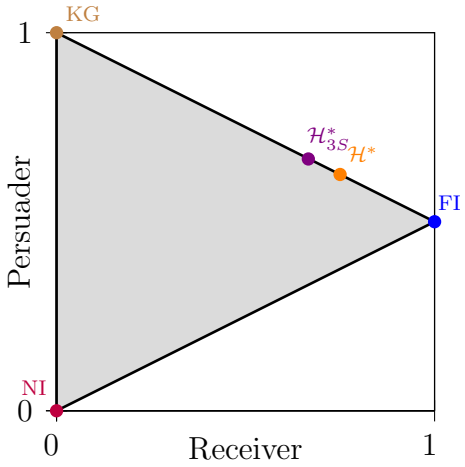

Figure 1: *Information structures in payoff space. Each player's payoff is expressed in percentage of her maximum attainable payoff. The grey triangle is the space of attainable payoffs, and the dots represent the payoffs achieved by different information structures.*

and 0 after signal $\underline{x}$. Indeed, in the absence of falsification, we have

$$\mathbb{E}_\tau(s|\overline{x}) = \overline{s}$$

and

$$\mathbb{E}_\tau(s|\underline{x}) \propto \pi_0(1-c)\overline{s} - (1-\pi_0)\underline{s} < 0,$$

so the strategy of the receiver is a best response to no falsification. Furthermore, deviating from no falsification to $\phi > 0$ would increase the probability that the receiver chooses approve by $(1-\pi_0)\phi c$, but also cost $(1-\pi_0)\phi c$. If $c = 0$, this test is just the uninformative test, but otherwise it gives valuable information to the receiver and prevents falsification.

I think we should just merge this with the two state case below

3 Model

As in the example, we consider a receiver who can choose between two actions labelled *approve* and *reject* for simplicity. The rejection payoff to the receiver is normalized to 0, whereas the approval payoff is given by the state of the world $s \in S \subseteq [-\underline{s}, \overline{s}]$, with $-\underline{s} < 0 < \overline{s}$, and $\{-\underline{s}, \overline{s}\} \subseteq S$. We refer to the case $S = \{-\underline{s}, \overline{s}\}$ as the binary state case. The incremental payoff of the agent from approval is positive and independent of the state of the world. We assume

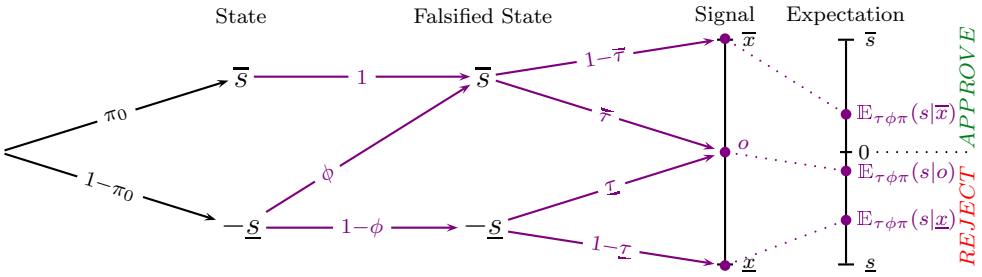

Figure 2: *A Better Test for the observable case. The expectation column shows how expectations associated with each signal shift under falsification.*

that the agent approves whenever she is indifferent.

Prior. The prior distribution for the state of the world has probability measure π, with full support on S. We denote its cdf F_π. We assume that $\mu_0 = \mathbb{E}_\pi(s) < 0$, so that the receiver-optimal action under the prior is to reject. We let $S^- = S \cap (-\infty, 0)$ and $S^+ = S \cap [0, \infty)$ denote the sets of negative and nonnegative states. We define

$$\mu_0^- = \mathbb{E}_\pi(s|s \in S^-) < \mu_0 < 0 < \mu_0^+ = \mathbb{E}_\pi(s|s \in S^+),$$

and let $\pi_0 = \pi(S^+)$ be the prior probability of nonnegative states. Finally, we let

$$s_0 = \max\{s' \in S : \mathbb{E}_\pi(s|s \geq s') \leq 0\}.$$

In particular, if π has no atom at s_0, then $\mathbb{E}_\pi(s|s \geq s_0) = 0$.

Timing. We first describe the timing of the game and then proceed to define tests, falsification and how the receiver's beliefs are formed.

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

$t \in S$ according to $\phi(s|t)$, and the test generates a public signal x about the falsified state of the world according to $\tau(x|t)$.

5. **Receiver decision:** The receiver forms beliefs and chooses to approve or reject.

Tests. A test is a Blackwell experiment (Blackwell, 1951, 1953): a measurable space of signals X, and a Markov kernel τ from S to $\Delta(X)$.

General Notations. The prior π and the test τ together define a joint probability measure on $X \times S$ that we simply denote by $\tau\pi$. Then, conditional on observing x, and in the absence of falsification, a receiver forms a belief about S that is given by the conditional probability measure which we denote by $\tau\pi_x$. Conditional on the state of the world s, and in the absence of falsification, the distribution of signals depends on τ only, and we denote it by τ_s.

Falsification. The agent can falsify the state of the world that is fed to the test. This is modeled as the choice of a falsification strategy ϕ which is a Markov kernel from S to ΔS. 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. The *truth-telling strategy* is the Markov kernel δ that maps each state s to the Dirac measure δ_s on S, which puts probability 1 on target state s when the true state is s.

Falsification comes at a cost given by a measurable non-negative real function $c : S \times S \to \mathbb{R}_+$, where we denote by $c(t|s)$ the cost of falsifying source state s as target state t. Together, the prior π and the falsification strategy ϕ define a joint probability measure that we denote $\phi\pi$ on $S \times S$. The cost of falsification strategy ϕ is then given by $C(\phi) = \int_{S\times S} c\, d\phi\pi$. The following are maintained assumptions about the cost function.

Assumption 1 (Maintained Cost Assumptions)**.** *We assume that truth telling is costless, that is $c(s|s) = 0$ for all $s \in S$. We also assume that c is either uniformly null or satisfies the following monotonicity properties: $c(t|s) < c(t'|s)$ for all s, t, t' such that $t' < t \leq s$ or $s \leq t < t'$; and $c(t|s) > c(t|s')$ for all s, s', t such that $s' < s \leq t$ or $t \leq s < s'$. Finally, we assume that the cost function is continuous in t.*

While the former are maintained assumptions, the following properties are not always assumed but are important for some results. The first two are familiar properties, and the last

one is a specific smoothness assumption that suits our purpose and that we call regularity.

Definition 1. *The cost function:*

(i) satisfies the triangular inequality if, for every triple of states $s, m, t \in S$,

$$c(t|m) + c(m|s) \geq c(t|s); \tag{TRI}$$

(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) 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 $t > s$,

$$c(t|s) \leq K(t - s).$$

When they exist, we denote the partial derivatives of the cost function by c_t and c_s. Next, we provide examples of cost functions that satisfy our assumptions.

Example 1 (Valid Cost functions)**.** All cost functions below satisfy all our maintained assumptions, as well as (TRI) and (UID) (proofs are available in APPENDIX REF).

1. **Monotonically Scaled Subadditive Cost:**

$$c(t|s) = \alpha(s) f(|t - s|) \mathbb{1}(t \geq s) + \frac{1}{\beta(s)} g(|t - s|) \mathbb{1}(t < s),$$

where $f, g : \mathbb{R}_+ \to \mathbb{R}_+$ are continuous, increasing and subadditive functions with $f(0) = g(0) = 0$, and $\alpha, \beta : S \to \mathbb{R}_+$ are strictly positive and nondecreasing. Furthermore, c is regular if f, g, α and β are continuously differentiable on their domain.

This class of cost functions includes the case where f and g are concave functions, and, in particular, the linear cost function $c(t|s) = |t - s|$. However, some convexity of f and g can also be accommodated if α and β increase sufficiently fast, as the next example illustrates.

2. **A Monotonically Scaled Convex Example:**

$$c(t|s) = \alpha e^{2\beta s}\left(|t-s| + \beta(t-s)^2\right)\mathbb{1}\left(t \geq s\right) + \alpha' e^{-2\beta' s}\left(|t-s| + \beta'(t-s)^2\right)\mathbb{1}(t < s),$$

with $\alpha, \beta, \alpha', \beta' > 0$, and where $\underline{s}$ and $\overline{s}$ satisfy...

Furthermore, c is regular.

$\diamond$

Posterior beliefs, actions and resulting payoffs. Together, the prior, the falsification strategy and the test define a joint distribution over $X \times S$ that we denote $\tau\phi\pi$. For each signal x occurring with positive probability according to $\tau\phi\pi$, the receiver forms a posterior belief in ΔX which we denote by $\tau\phi\pi_x$. Let $\mu(x|\tau,\phi) = \int_S s\, d\tau\phi\pi_x(s)$ denote the expected state according to this belief. The receiver approves whenever $\mu(x|\tau,\phi) \geq 0$. We let $\bar{X}(\tau,\phi) = \{x : \mu(x|\tau,\phi) \geq 0\}$ denote the approval set of the receiver. Then, the ex ante probability of approval is given by

$$A(\tau,\phi) = \int_{\bar{X}(\tau,\phi)\times S} d\tau\phi\pi,$$

and the agent's payoff by

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

whereas the receiver's payoff is given by the expected state conditional on approval

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

Solution Concept and Equilibrium Definition. Our equilibrium concept is perfect Bayesian equilibrium. In equilibrium falsification is correctly anticipated. However, deviations from a falsification strategy can go undetected since by construction any signal x can arise from *some* state and the latter is unknown.

Non-committed versus committed falsification Typically falsification is not observed and deviations are not detectable. We call this the *unobservable* case (and sometimes "non-committed" to contrast it with the case that follows). When the agent's deviations are undetected the receiver's belief after each signal remains unchanged and the ex-ante approval

probability resulting from a deviation to ϕ' when the receiver anticipates ϕ therefore yields the approval probability:

$$A(\tau, \phi, \phi') = \int_{\bar{X}(\tau,\phi)\times S} d\tau\phi'\pi.$$

Then, given a test τ, there exists an equilibrium with falsification strategy ϕ if and only if, for every falsification strategy ϕ',

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

The *interim* probability that state s is approved given τ, ϕ is: $a(s; \tau, \phi) \equiv \int_{\bar{X}(\tau,\phi)} d\tau\phi$.

If falsification is *observable* deviations are detected.[9] This case can also be casted as commitment to a particular 'manipulation' technology: The agent is akin or to a "constrained information designer" and can only induce information structures that are feasible given the (exogenous) test in place and his falsification capabilities. When deviations are observable and the agent deviates from ϕ, to ϕ' each signal realization x changes "meaning," from $\mu(x|\tau, \phi)$ to $\mu(x|\tau, \phi')$ and the set of "aprove" signals becomes $\bar{X}(\tau, \phi')$. In this case, given a test τ, there exists an equilibrium with falsification strategy ϕ if and only if, for every falsification strategy ϕ',

$$U(\tau, \phi) \geq U(\tau, \phi').$$

Both benchmarks are relevant depending on the situation. We derive receiver-optimal tests for both the observable and unobservable falsification settings and compare their features and payoffs they yield to the agent and the receiver.

Receiver-optimal tests We are mainly interested in optimizing the test for the receiver when the agent can falsify. When falsification is unobservable, a receiver optimal test solves:

$$\begin{aligned} &\sup_{\tau,\phi} V(\tau, \phi) \\ &\text{s.t. } U(\tau, \phi) \geq A(\tau, \phi, \phi') - C(\phi'), \ \forall \phi', \end{aligned}$$

[9] They are also detectable when there is a very large number of items and the receiver can infer the actual degree of falsification from the empirical distribution of results. Details of how detection works can be found in Perez-Richet and Skreta (2018).

and when falsification is observable, a receiver optimal test solves:

$$\sup_{\tau,\phi} V(\tau,\phi)$$
$$\text{s.t. } U(\tau,\phi) \geq U(\tau,\phi'),\ \forall \phi'.$$

Note that we consider ex ante falsification–i.e. the agent chooses ϕ before observing the state. For the case of unobservable falsification, there exist a receiver optimal test for which the ante and interim falsification choices *coincide* (so imposing the more demanding interim falsification constraint does not reduce the value of the program, as we establish in Proposition 2). For committed falsification our analysis extends to interim manipulations by the agent with small modifications. Details are available upon request.

These programs are intricate because, the ultimate information structure is a composition of the test τ and the falsification strategy ϕ. It would be helpful to have a revelation-principle type of result that would allow us to restrict attention to truth-telling strategy, that is falsification strategies that map each state s to the Dirac measure δ_s. Under costly falsification, there is no general falsification proofness principle akin to the revelation principle in mechanism design. In fact, we show that, in general, the receiver optimal test *requires* falsification by positive states. However, there are some important cases in which the principle holds: when falsification is costless, or when the state space is binary:

Proposition 1 (Costless falsification: Falsification Proofness Principle)**.** *If falsification is costless or the state space S is binary, then, for every test τ such that ϕ is an equilibrium falsification strategy, there exists a test τ' under which truth-telling δ is an equilibrium and $U(\tau,\phi) = U(\tau',\delta)$ and $V(\tau,\phi) = V(\tau',\delta)$. This is true regardless of whether falsification is observable or unobservable.*

Pareto frontier, Bergemann Morris perspective etc.

4 Unobservable Falsification

4.1 Preliminary Results

Recommendation Principle. When falsification is unobservable, the falsification strategy is analogous to a costly reporting strategy. Mimicking standard results as those in Myerson (1982) and Kamenica and Gentzkow (2011), we can establish a recommendation principle according to which signal realizations of any test can be garbled into two signal realizations that are action recommendations (so in our case "approve" or "reject") without changing equilibrium falsification strategy, payoffs and interim approval probabilities. The proof of this result can be found in Appendix A.

Lemma 1 (Recommendation Principle)**.** *Let ϕ be an equilibrium falsification strategy under τ. Then the test τ' with binary signal space $X' = \{approve, reject\}$ defined by*

$$\tau'(approve|s) = \tau(\bar{X}(\tau, \phi)|s)$$

is such that ϕ is an equilibrium under τ' and

1. *Receiver follows recommendation: $\bar{X}(\tau', \phi) = \{Pass\}$,*
2. *Equilibrium interim approval probabilities are unchanged: $a(s; \tau, \phi) = a(s; \tau', \phi)$,*
3. *Equilibrium payoffs are unchanged: $U(\tau, \phi) = U(\tau, \phi')$ and $V(\tau, \phi) = V(\tau', \phi)$.*

Note that, as is usual in this type of results, by definition the new test yields the same interim probability of approval for fixed measure of states fed to it: $\phi\pi$. However, in our setting, in addition to making sure the receiver follows the "new" recommendation, we have to make sure that the new test does not yield a new falsification strategy as a best response. 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.

Lemma 1 allows us to restrict our discussion to binary tests such that the receiver follows recommendations. Therefore, for the remainder of our analysis of unobservable falsification, we redefine tests as measurable functions $\tau : S \to [0, 1]$, where $\tau(s)$ is the *nominal passing probability* of state s. Falsification may of course induce a true passing probability that differs

from the nominal one. A receiver optimal test solves the following program:

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s) \tag{EP}$$

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

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

where the constraint, (EOF) is the *ex-ante* optimal falsification constraint for the agent, and the (RO) constraint it the receiver's obedience constraint.

First, note that the left-hand side term in the obedience constraint is equal to the expected payoff of the receiver, which is also the objective function. Since the uninformative test that always recommends fail yields a null payoff for the receiver and satisfies both constraints for any falsification strategy, it is clear that the obedience constraint is redundant and can be omitted from the program. A corollary of this remark is that, under unobservable falsification, the receiver does not benefit from commitment to an approval strategy, as the program of such a receiver is exactly the one above without the obedience constraint.

A second remark is that the value of the program when we impose the ex ante optimal falsification constraint is identical to the value of that with the interim one (IOF):

$$\sup_{\tau,\phi} \int_{S\times S} s\tau(t)d\phi\pi(t,s) \tag{IP}$$

$$\text{s.t. } \phi\left(\Phi(s;\tau)|s\right) = 1, \quad \forall s \in S \tag{IOF}$$

$$\text{where } \Phi(s;\tau) = \text{argmax}_t \ \tau(t) - c(t|s) \tag{optimal falsification targets}$$

Proposition 2. *When falsification is unobservable, a receiver-optimal test yields the same payoff to the receiver, regardless of whether the agent chooses his falsification strategy ex-ante or at the interim stage. That is, the value of program* (EP) *is equal to that of* (IP)*.*

In what follows, we thus focus on Program (IP). This program is analogous to that 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 premium from getting the good; the principal can commit to a probabilistic allocation rule τ contingent on reported state, misreporting has

a cost given by c. As mentioned in the introduction, there is a relatively small literature studying allocation problems without transfers. Our problem differs because misreporting is costly. This, on one hand makes lies harder, on the other, restricting attention to tests that induce truth-telling is with loss of generality.

These types of problems where transfers are not available are studied under various conditions in the literature. REVIEW HERE BRIEFLY.

There is, however, a Falsification Proofness Principle when falsification is costless, as we have established in Proposition 1. And, is this case it provides a direct solution to our problem: nothing can be achieved. Indeed, we can then restrict attention to obedient pass-fail tests that are falsification proof. If such a test recommends to pass some state with positive probability, then it is optimal for the agent to falsify every other state with lower passing probability as this state. Hence the only falsification-proof tests recommend all states to pass with uniform probability. But such a test conveys no information.

we are using approve/ pass/ action 1 interchangeably–I rewrote action 1 but we have to settle.

Corollary 1. *If falsification is costless and the agent cannot commit, no information can be conveyed in equilibrium, and* $U^* = V^* = 0$.

4.2 The Binary State Case

In the binary state case, the recommendation principle makes it particularly easy to characterize the set of feasible tests and payoffs. Indeed, by Lemma 1, we can focus on tests with only a passing and a failing signal. Hence, we can describe a test as a pair of "aproval" probabilities $\tau = (\overline{\tau}, \underline{\tau})$ that satisfy the obedience constraint $\overline{\tau}\pi_0\overline{s} - \underline{\tau}(1-\pi_0)\underline{s}$. Then, we can also easily obtain a trivial falsification-proofness: If $\overline{\tau} - \underline{\tau} > \underline{c}$ then $-\underline{s}$ falsifies as $\overline{s}$ and no state passes, so the probability of passing $\overline{s}$ is bounded above $\underline{c} + \underline{\tau}$ and without loss we can focus on tests that satisfy $\overline{\tau} - \underline{\tau} \leq \underline{c}$, where $\underline{c}$ denotes the cost of falsifying the low type as the low type.[10] For such tests, the indifference curves of the receiver are described by the equation $U = \overline{\tau}\pi_0\overline{s} - \underline{\tau}(1-\pi_0)\underline{s}$, and the indifference curves of the agent by $V = \pi_0\overline{\tau} + (1-\pi_0)\underline{\tau}$. Then, elementary algebra yields the following characterization:

[10] Note that the cost of falsifying the high type as the low type plays no role in this case.

Proposition 3. *With a binary state space, the set of falsification proof and obedient tests is given by*

$$\mathcal{T} = \begin{cases} \text{co}\{\tau_A, \tau_R, \tau_0\} & \text{if } \underline{c} \leq 1 - \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}} \\ \text{co}\{\tau_A, \tau_R, \tau_0, \tau_P\} & \text{if } \underline{c} \geq 1 - \frac{\pi_0 \overline{s}}{(1-\pi_0)\underline{s}} \end{cases},$$

where, $k = \min(\underline{c}, 1)$, $\tau_A = \big(\min\big(-k\pi_0\overline{s}/\mu_0, \pi_0\overline{s}/(1-\pi_0)\underline{s}\big), \min\big(k - k\pi_0\overline{s}/\mu_0, 1\big)\big)$ *is the agent-optimal test,* $\tau_R = \big(0, k\big)$ *is the receiver-optimal test,* $\tau_0 = (0,0)$ *is the uninformative test which is pessimal for both players, and* $\tau_P = (1-k, 1)$ *is feasible only in the high cost case, and is then the preferred test of a planner with equal weights on both players. The corresponding feasible payoff set is given by*

$$\mathcal{U} = \begin{cases} \text{co}\{\mathfrak{u}_A, \mathfrak{u}_R, \mathfrak{u}_0\} & \text{if } \underline{c} \leq 1 - \frac{\pi_0 \overline{s}}{1-\pi_0)\underline{s}} \\ \text{co}\{\mathfrak{u}_A, \mathfrak{u}_R, \mathfrak{u}_0, \mathfrak{u}_P\} & \text{if } \underline{c} \geq 1 - \frac{\pi_0 \overline{s}}{1-\pi_0)\underline{s}} \end{cases},$$

where $\mathfrak{u}_A = \big(0, \min(-k\pi_0\overline{s}/\mu_0 + 1 - \pi_0, \pi_0(\overline{s} + \underline{s})/\underline{s})\big)$, $\mathfrak{u}_R = \big(k\pi_0\overline{s}, k\pi_0\big)$, $\mathfrak{u}_0 = (0,0)$, *and, in the high cost case,* $\mathfrak{u}_P = \big(\mu_0 - k(1-\pi_0)\underline{s}, 1 - (1-\pi_0)k\big)$.

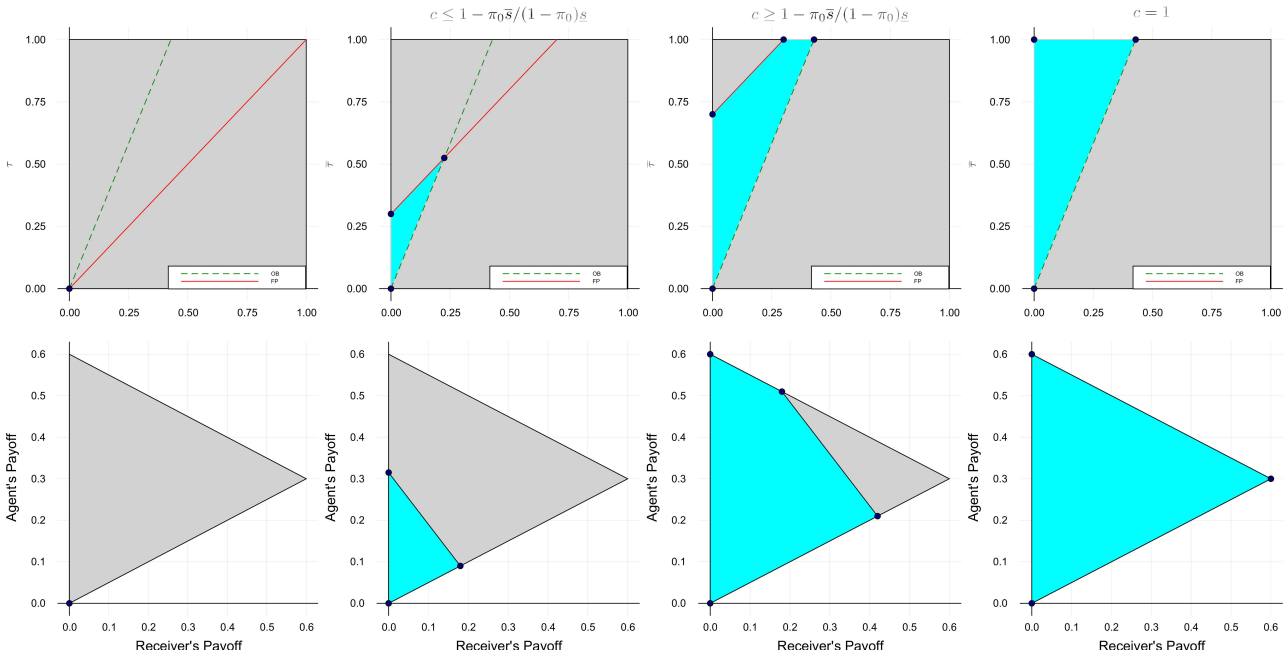

Figure 3

4.3 Continuus State: Space A Receiver Optimal Test

Next, we find a receiver optimal test when S is the full interval $[-\underline{s}, \overline{s}]$, and falsification is costly. Our characterization requires that the cost function satisfies the triangular inequality. Using Proposition 2, we proceed by solving (IP). Our first result is a weak and partial Falsification Proofness Principle. In the usual revelation principle, the idea is to give a falsifying agent directly the allocation and payoff he gets by falsifying, thus not modifying the payoff of the principal. In the absence of transfers, it is impossible to maintain both allocation and payoff, hence we only give the falsifying agent the same payoff. But, we make sure to do so in a way that favors the receiver (which is our equivalent of a principal). This is the sense in which the falsification proofness principle of this result is weak. It is partial because this operation only works for negative types. The role of the triangular inequality for the revelation principle with costly misreporting and transferable utilities is pointed out in Kephart and Conitzer (2016). Because we do not have transfers, our proof differs, but relies on a similar intuition.

Lemma 2. *If the cost function satisfies* (TRI)*, then, for every test τ and incentive compatible falsification strategy ϕ, there exists a test $\tilde{\tau}$ and an incentive compatible falsification strategy $\tilde{\phi}$, such that $\tilde{\phi}(s) = \delta_s \mathbb{1}(s < 0) + \phi(s) \mathbb{1}(s \geq 0)$, and no player is worse off: $U(\tilde{\tau}, \tilde{\phi}) \geq U(\tau, \phi)$ and $V(\tilde{\tau}, \tilde{\phi}) \geq V(\tau, \phi)$.*

Remark 1. The role of the triangular inequality property of costs in the proof of Lemma 2 is analogous to that in the revelation principle in Kephart and Conitzer (2016) for the FUVT–"fixed utilities" variable transfers case." Here we are keeping the sender's utility fixed but varying the social choice function: When we replace the original measure $\tau\phi\pi$ with $\tau^{FP}\pi$ the agent saves on falsification costs. To keep the payoff the same, τ^{FP} has a smaller probability of passing. This benefits the receiver for negative states but is worse for positive states and the overall effect is ambiguous. Thus in our setting, because we don't have transfers, we don't have enough flexibility to design a test that is at the same time falsification proof and leaves both the agent's and the receiver's payoffs unchanged. Indeed, the result is false if we insist on falsification-proofness also for the positive states and as we establish next the optimal test involves falsification by the positive states.

Lemma 2 implies that we can restrict attention to tests that are falsification-proof for negative states. Next, we show that we can further restrict attention to tests such that positive

states do not have any incentive to falsify as negative states.[11]

Lemma 3. *Let τ be a test and ϕ an optimal falsification strategy, such that, for every $s < 0$, $\phi(s) = \delta_s$. Then there exists a test $\tilde{\tau}$ and an optimal falsification strategy $\tilde{\phi}$ such that, for every $s < 0$, $\tilde{\phi}(s) = \delta_s$, and, for every $s \geq 0$, $\phi\big(S^-|s\big)\big) = 0$, and, furthermore, no player is worse off: $U(\tilde{\tau}, \tilde{\phi}) \geq U(\tau, \phi)$ and $V(\tilde{\tau}, \tilde{\phi}) \geq V(\tau, \phi)$.*

An optimal class of tests We now consider a class of tests defined 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. Let $\check{s}(p, \hat{s}) = \min\{s \leq \hat{s} : c(\hat{s}|s) \leq p\}$, and define the lower, middle and upper intervals $I_L = \big[-\underline{s}, \check{s}(p, \hat{s})\big)$, $I_M = \big[\check{s}(p, \hat{s}), \hat{s}\big]$ and $I_U = \big(\hat{s}, \overline{s}\big]$. Our class consists of the tests defined by

$$\tau_{p,\hat{s}}(s) = p\,\mathbb{1}(s \in I_U) + \big(p - c(\hat{s}|s)\big)\,\mathbb{1}(s \in I_M),$$

and such that $\check{s}(p, \hat{s}) \leq 0$.

These tests satisfy a number of interesting properties: First, the nominal passing probability is continuous, strictly increasing on I_M, and constant on each of the remaining intervals. Second, truth-telling is an optimal falsification strategy if (TRI) holds. However, truth-telling is not receiver-preferred optimal falsification strategy. Indeed, third, all states in I_M are indifferent between truth-telling and falsifying as $\hat{s}$. The receiver-preferred equilibrium feasible falsification strategy is that all such states do falsify as $\hat{s}$ if positive, and not falsify if negative. As a consequence, all positive states pass with probability p, whereas negative states pass with their nominal probability. These results are summarized in the following lemma:

Lemma 4. *For every $p \in [0, 1]$, and every $\hat{s} \in S^+$, the test $\tau_{p,\hat{s}}$ satisfies the following properties:*

(i) $\tau_{p,\hat{s}}$ is continuous on S, strictly increasing on I_M and constant on each of the intervals I_L and I_U.

(ii) If the cost function satisfies (TRI), then $s \in \Phi(s; \tau_{p,\hat{s}})$ for every $s \in S$.

(iii) For every $s \in I_M$, $\hat{s} \in \Phi(s; \tau_{p,\hat{s}})$.

We show that, when looking for a receiver optimal test, we can restrict attention to tests within this class.

[11] This is the step of the proof that requires S to be the full interval.

Proposition 4. *Suppose the cost function satisfies* (TRI)*. Then, for every test* τ*, and every incentive compatible falsification strategy* ϕ*, there exists a test* $\tau_{p,\hat{s}}$ *such that the falsification strategy*

$$\phi_{p,\hat{s}}(s) = \delta_s \mathbb{1}\left(s \in S^- \cup I_U\right) + \delta_{\hat{s}} \mathbb{1}\left(s \in S^+ \cap I_M\right)$$

is optimal, and $V\left(\tau_{p,\hat{s}}, \phi_{p,\hat{s}}\right) \geq V(\tau, \phi)$*.*

Finally, we characterize the receiver optimal test within our class, and use the steps above to conclude that it is a receiver optimal test among all tests.

Theorem 1. *Suppose the cost function satisfies* (TRI)*. Then* $(\tau_{p^*,s^*}, \phi_{p^*,s^*})$ *maximizes* (EP)*, where*

$$p^* = \min\left\{c(\overline{s}|s_0), 1\right\},$$

and

$$s^* = \max\left\{s \in S : c(s|0) \leq 1\right\}.$$

Furthermore, the receiver gets her first-best payoff if and only if $c(\overline{s}|0) \geq 1$*. However, the pair of resulting payoffs* (U^*, V^*) *never lies on the Pareto frontier.*

With this optimal test, types in $S^+ \cap I_M$ are indifferent between falsifying as s^* and truth-telling, and are made to falsify only for the benefit of the receiver. It is easy to find other tests that achieve the same receiver payoffs and strengthen the incentive of types in $S^+ \cap I_M$ to falsify by lowering their nominal passing probability. For example, consider any test τ such that $\tau(s) = \tau_{p^*,s^*}(s)$ for every $s \in S^- \cup I_U$, and $\tau(s) \leq \tau_{p^*,s^*}(s)$ for $s \in S^+ \cap I_M$. It is incentive compatible for all states in $S^+ \cap I_M$ to falsify as s^*, and for all other states not to falsify. Both tests generate the same information structure under this falsification strategy, and receiver and agent get the same payoffs.

4.4 An Optimal Falsification Proof Test

Cheating may have a negative effect on society as illustrated in Galbiati and Zanella (2012); Ajzenman (2018); Alm et al. (2017); Rincke and Traxler (2011), among many others. Because of this negative externality, a test designer may want to employ only tests that do not generate falsification. Motivated by this, we now characterize the receiver-optimal falsification proof test. In this section, we make the following additional assumptions

Assumption 2. *$c(t|s)$ is regular in the sense of Definition 1 and satisfies*(UID) *and* (TRI)*. Furthermore, the state space is the whole interval $S = [-\underline{s}, \overline{s}]$, and the prior is atomless.*

Building on Proposition 2, we can write the corresponding program as follows[12]:

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

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

We start by showing that we can restrict attention to K-Lipschitz and nondecreasing test functions. Indeed, continuity is implied by (FPIC). For monotonicity, we show that replacing a falsification proof test by the highest monotonic function below it for negatives states, and the highest monotonic function above it for nonnegative states generates a test that is monotonic, yields more favorable passing probabilities for the receiver, and preserves falsification proofness.

Lemma 5. *Suppose that c is regular. Let τ be a test that satisfies* (FPIC)*. Then τ is continuous, and there exists a K-Lipschitz and nondecreasing test function $\hat{\tau}$ that also satisfies* (FPIC) *and makes the receiver better off.*

Lipschitz continuity implies that we can now focus attention on tests that are almost everywhere differentiable with derivative τ' bounded in $[0, K]$, 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 $\{\tau'(s)\}_{s \in S}$. Then, we can use integration by part to rewrite the objective function in (FPProg) as

$$\underline{\tau}\mu_0 + \int_S \tau'(z)J(z)dz,$$

where $J : S \to \mathbb{R}$

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

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

[12]Strictly speaking, we first need to make sure that there is a falsification-proof test that gives the receiver an expected payoff above 0 to ensure that the obedience constraint is redundant, but the uninformative test clearly achieves that.

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 $s < t$, $\int_s^t \tau'(z)dz \leq c(t|s)$. Reducing $\underline{\tau}$ increases the objective function as $\mu_0 < 0$, relaxes the probability constraint, and has no effect on the incentive constraints, implying that it is optimal to set $\underline{\tau} = 0$.

Next, we treat the probability constraint with the Lagrangian method, and therefore associate it with a Lagrange multiplier $\lambda \geq 0$ and rewrite the objective function accordingly which yields the Lagrangian problem:

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

The Lagrangian problem is to maximize $\mathcal{L}(\tau', \lambda)$ where $\tau' : S \to [0, K]$ is feasible if, for every $s < t$, $\int_s^t \tau'(z)dz \leq c(t|s)$, 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 $[s_*, s^*]$ such that $J(s_*) = J(s^*) = \lambda$.

The following result combines these observations with a version of the Lagrangian sufficiency theorem.

Lemma 6. *Suppose that there exists $\hat{\lambda} \geq 0$, and a feasible $\hat{\tau}'$ such that:*

(a) $\hat{\lambda} = 0$ or $\int_S \tau'(z)dz = 1$;

(b) For every feasible τ', $\mathcal{L}(\hat{\tau}', \hat{\lambda}) \geq \mathcal{L}(\tau', \hat{\lambda})$.

Then there exists an interval $[s_*, s^*]$ such that:

(i) $s_0 \leq s_* \leq 0 \leq s^* \leq \overline{s}$ and $J(s_*) = J(s^*) = \hat{\lambda}$;

(ii) $\hat{\tau}'(s) = 0$ for every $s \notin [s_*, s^*]$;

(iii) The test $\hat{\tau}(s) = \left(\int_{s_*}^{s^* \wedge s} \hat{\tau}'(z)dz\right) \mathbb{1}(s \geq s_*)$ is a falsification-proof receiver optimal test.

First, note that each choice of $s_* \in [s_0, 0]$ uniquely pins down $s^* = m(s_*)$, 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 $\int_{s_*}^{m(s^*)} s dF_\pi(s)$. In particular, note that s_0 is matched with $m(s_0) = \overline{s}$. This matching function will play an important role in the characterization of the optimal test.

Following the Lagrangian method, we next choose a value for the Lagrange multiplier, which is equivalent to choosing s_* by the first point of Lemma 6. We choose

$$s_* = \min\{s \in [s_0, 0] : c(m(s)|s) \leq 1\}, \tag{1}$$

so that $s_* = s_0$ whenever $c(\overline{s}|s_0) \leq 1$. Let $\lambda^* = J(s_*)$ be the corresponding Lagrange multiplier.

Instead of solving the Lagrangian problem, we go back to the original program, but focusing on tests τ that are constant outside of $[s_*, s^*]$, and such that $\tau(s_*) = 0$. However, we also relax the program by getting rid of the constraint that $\tau(s^*) \leq 1$, and only keeping the incentive constraints for pairs (s, t) such that $s_* \leq s \leq 0 \leq t < s^*$. 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 $\mu^* = \int_0^{s^*} s dF_\pi(s)$, the objective function of the program becomes

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

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

To summarize, the new relaxed and reformulated program is

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

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

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

where $\mathcal{M}(P, Q)$ is the set of joint distributions on $Z \times Y$ with marginals Q on Z, and P on Y. By (UID), the transportation cost function of this problem, $c(z| - y)$ is submodular, implying

a well known solution for both problems[13]. Rewriting this solution[14] in terms of our initial notations, and completing for states outside of $[s_*, s^*]$, we obtain the test

$$\tau^*(s) = \left(-\int_{s_*}^{s} c_s\big(m(x)|x\big)dx \right) \mathbb{1}\left(s \in [s_*, 0]\right) + \left(c(s^*|s_*) - \int_{s}^{s^*} c_t\big(x|m^{-1}(x)dx\big) \right) \mathbb{1}\left(s \in (0, s^*]\right) + \mathbb{1}\left(s > s^*\right)$$

The folllowing theorem shows that τ^* solves our initial problem.

Theorem 2. *The test τ^* solves (FPProg) and is therefore a receiver-optimal falsification-proof test. The corresponding receiver's payoff is given by*

$$U(\tau^*, \delta) = \int_{s_*}^{0} -sc\big(m(s)|s\big)dF_\pi(s) = \int_{0}^{s^*} tc\big(t|m^{-1}(t)\big)dF_\pi(t).$$

Furthermore, the outcome (τ^*, δ) is Pareto inefficient.

Proof of Theorem 2. need to make sure that test satisfies all constraints, then go back the chain to apply Lemma 6. □

To understand the logic of the proof, note that, either $s_* = s_0$, and then $\lambda^* = 0$, and $\tau^*(s^*) = c(\overline{s}|s_0) \leq 1$, so that τ^* satisfies the probability constraint, or $s_* > s_0$, and then $\lambda^* > 0$ and $\tau(s^*) = c(s^*|s_*) = 1$ so the probability constraint is satisfied with equality. Hence, τ^* satisfies the probability constraint, and, in addition, point (a) of Lemma 6 holds. To show point (b), note that, by point (ii) of the same lemma, optimizing the Lagrangian is equivalent to solving our initial problem restricted to tests that are constant outside of $[s_*, s^*]$. The dual Monge-Kantorovich problem we obtained is a relaxed version of that, so we need to verify that τ^* satisfies the omitted incentive constraints. In the proof, we show that this is ensured by (TRI). Then Lemma 6 allows us to conclude.

The payoff characterization is a consequence of duality. Note that the solution of the primal Monge-Kantorovich 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. We show in SECTION REFF that the primal Monge-Kantorovich problem has a nice interpretation as it corresponds to the optimal

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

[14] 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 $\tau^*(s_*) = 0$.

falsification problem of a committed agent facing a fully revealing test, with a modified cost function.

5 Falsification with Commitment

We assume upward only falsification.

5.1 Preliminary Results

Falsification Proofness Principle. With committed falsification, we have a Falsification Proofness Principle that exists under the same exact conditions as in the uncommitted case

Normalization by the mean.

Commitment or Not. We show that any feasible equilibrium outcome without commitment can be replicated with commitment. More precisely

Proposition 5. *Any uncommitted equilibrium falsification strategy ϕ under test τ that is upward, is also an uncommitted upward only equilibrium falsification strategy under τ. Any uncommitted upward only equilibrium falsification strategy under τ is also a committed (upward only) equilibrium falsification strategy under τ.*

Proof. The first implication is obvious as the agent simply has less available deviations when feasible falsification strategies are required to be upward. For the second implication, we note that, by the recommendation principle, we can □

In particular, since the receiver optimal tests (falsification-proof or not) we characterized in SECTION REFERENCE lead to an equilibrium falsification strategy that is upward, we have the following corollary.

Corollary 2. *With receiver-optimal testing or receiver-optimal falsification-proof testing, the receiver is better off under committed falsification.*

5.2 Fully Revealing Tests: Interpreting Duality

5.3 The Binary State Case

In the binary state case, it is possible to characterize an optimal test in closed form. Our characterization relies on the falsification-proofness principle, and on the representation of tests as the distribution of expectations they generate for the receiver, which amounts to relabelling signals as expectations and possibly involves some garbling, as signals with the same associated expectation are merged. Lemma 7 (WILL BE MOVED TO APPENDIX – this can probably be proved beyond the binary case) shows that this is without loss of generality. It is well known that this representation is without loss of generality in the absence of falsification, so the key is to show that it does not affect the set of optimal falsification strategies, nor their payoff for the agent.

For a test τ, let H_τ be the corresponding cdf of associated expectations on $[-\underline{s}, \overline{s}]$, that is

$$H_\tau(y) = \tau\big(\{x \in X : \mathbb{E}_{\tau\pi}(s|x) \leq y\} \times S\big).$$

The martingale property implies that $\int_{-\underline{s}}^{\overline{s}} x dH_\tau(x) = \mu_0$, or equivalently, integrating by part, that

$$\int_{-\underline{s}}^{\overline{s}} H_\tau(x)dx = \overline{s} - \mu_0. \tag{MP}$$

In the other direction, to any cdf H on $[-\underline{s}, \overline{s}]$ that satisfies (MP), we can associate a unique test τ_H^E with signal space $X^E = [-\underline{s}, \overline{s}]$ such that $H_{\tau_H^E} = H$. Note that uniqueness is satisfied only in the binary case. To see this, first define the function $\mathcal{H}(x) = \int_{-\underline{s}}^{x} H(y)dy$ from $[-\underline{s}, \overline{s}]$ to $[0, \overline{s} - \mu_0]$, which is nondecreasing and convex, with $\mathcal{H}(0) = 0$ and $\mathcal{H}(\overline{s}) = \overline{s} - \mu_0$. Then, define τ_H^E by

$$\tau_H^E\big(\{x \in X : \mathbb{E}_{\tau\pi}(s|x) \leq y\}|\overline{s}\big) = \frac{1}{\mu_0 + \underline{s}}\big\{(y + \underline{s})H(y) - \mathcal{H}(y)\big\} = \overline{H}(y)$$

and

$$\tau_H^E\big(\{x \in X : \mathbb{E}_{\tau\pi}(s|x) \leq y\}| - \underline{s}\big) = \frac{1}{\overline{s} - \mu_0}\big\{(\overline{s} - y)H(y) + \mathcal{H}(y)\big\} = \underline{H}(y),$$

so that $\overline{H}$ and $\underline{H}$ are the cdfs of signals respectively generated by the high and low states $\overline{s}$ and $-\underline{s}$.

Lemma 7 (Normalization by the Mean)**.** *Let τ be a falsification-proof test. Then $\tau_{H_\tau}^E$ is also*

falsification-proof and generates the same payoffs, that is $U(\tau,\delta) = U(\tau^E_{H_\tau},\delta)$, *and* $V(\tau,\delta) = V(\tau^E_{H_\tau},\delta)$. *Furthermore, it satisfies the normalization by the mean property* $\mathbb{E}_{\tau^E_{H_\tau}\pi}(s|x) = x$.

As a consequence, we can restrict attention to tests of the form τ^E_H, where H is a cdf on X^E that satisfies the martingale property. For any such cdf H, we define the function

$$H_\ell(x) = \lim_{\substack{y\to x\\ y<x}} H(y),$$

which is also the left derivative of $\mathcal{H}$ at x, and gives the probability of the set $[0,x)$ under H.

In the absence of falsification, the payoffs obtained by the receiver and the agent under $\mathcal{H}$ are respectively given by[15]

$$U(\tau^E_H,\delta) = \int_0^{\overline{s}} x dH(x) = \mu_0 + \mathcal{H}(0),$$

and

$$V(\tau_H,\delta) = 1 - H_\ell(0).$$

Next, we consider the effect of falsification on the receiver. Here, we use $\phi \in [0,1]$ to denote the probability that the low state falsifies as the high state. For any $\phi \in (0,1)$, falsification pushes the low state towards nonegative signals $x \geq 0$ at a higher rate than in the absence of falsification, thus lowering the expectation formed by the receiver when observing x, and resulting in the receiver no longer approving following some nonnegative signals. In fact, falsification results in a new threshold signal $\hat{x}(\phi)$ such that the receiver only approves for signals $x \geq \hat{x}(\phi)$. Interestingly, this threshold is independent of the test.

Lemma 8. *With falsification, there exists a threshold* $\hat{x}(\phi)$ *such that the receiver approves for signals* $x \geq \hat{x}(\phi)$, *and rejects otherwise, where*

$$\hat{x}(\phi) = \frac{-\mu_0 \underline{s}\phi}{\pi_0(\overline{s}+\underline{s}) - \phi\underline{s}}\,\mathbb{1}\left(\phi \leq \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}\right) + \overline{s}\,\mathbb{1}\left(\phi > \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}\right)$$

is continuous, increasing in ϕ *on* $[0, \pi_0\overline{s}/(1-\pi_0)\underline{s}]$, *constant elsewhere, and ranges from 0 to* $\overline{s}$.

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

This result has simple but very useful implications for the formulation of the optimal falsification problem of the agent.

Corollary 3. *Falsification levels outside of* $[0, \pi_0\overline{s}/(1-\pi_0)\underline{s}]$ *are dominated. Furthermore there is a one-to-one relationship between any* ϕ *in this range and the threshold if generates, implying that we can reformulate the receiver's optimal falsification problem as the choice of a threshold* $x \in [0, \overline{s}]$*, which induces a falsification level*

$$\hat{\phi}(x) = \frac{(\underline{s}+\mu_0)x}{(x-\mu_0)\underline{s}}.$$

The agent's payoff is then given by

$$V\left(\tau_H, \hat{\phi}(x)\right) = 1 - \left(1 + \frac{x}{\underline{s}}\right) H_\ell(x) + \frac{x}{\underline{s}(x-\mu_0)}\mathcal{H}(x) - \frac{(1-\pi_0)(\underline{s}+\mu_0)x}{(x-\mu_0)\underline{s}}c.$$

Using the Falsification Proofness Principle, we can now reformulate the program for finding a receiver-optimal test as that of choosing a test function $\mathcal{H}$ to maximize $\mathcal{H}(0)$, under the falsification proofness constraint that, for every $x \in [0, \overline{s}]$,

$$H_\ell(x) - \frac{x}{(\underline{s}+x)(x-\mu_0)}\mathcal{H}(x) \geq \frac{\underline{s}}{\underline{s}+x}H_\ell(0) - \frac{\gamma x}{(x-\mu_0)(\underline{s}+x)}, \tag{CFPIC}$$

where $\gamma = (\overline{s}-\mu_0)(\underline{s}+\mu_0)c/(\underline{s}+\overline{s})$.

A first remark is that we can focus on test functions that are linear on $[-\underline{s}, 0]$. Indeed, for any test function $\mathcal{H}$ that satisfies (CFPIC), the test function

$$\tilde{\mathcal{H}}(x) = \frac{\mathcal{H}(0)}{\underline{s}}(x+\underline{s})\,\mathbb{1}(x \leq 0) + \mathcal{H}(x)\,\mathbb{1}(x > 0)$$

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

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 ϕ. Denoting by κ

its slope below 0, this test function must solve the indifference differential equation[16]

$$H(x) - \frac{x}{(\underline{s}+x)(x-\mu_0)}\mathcal{H}(x) = \frac{\kappa \underline{s}}{\underline{s}+x} - \frac{\gamma x}{(x-\mu_0)(\underline{s}+x)} \tag{IDE}$$

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 $\mathcal{H}(\overline{s}) = \overline{s} - \mu_0$ which pins down the κ to a value that we denote by κ^*_γ yielding a unique test function

$$\mathcal{H}^*_\gamma(x) = \kappa^*_\gamma(x+\underline{s}) + \left(\kappa^*_\gamma(\mu_0+\underline{s}) - \gamma\right)\left\{\left(\frac{x-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}} - 1\right\}\mathbb{1}(x>0),$$

where

$$\kappa^*_\gamma = \frac{\overline{s}-\mu_0+\gamma\left\{\left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}} - 1\right\}}{\overline{s}-\mu_0+(\underline{s}+\mu_0)\left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}}.$$

We show that $\mathcal{H}^*_\gamma$ is also receiver-optimal.

Theorem 3. *$\mathcal{H}^*_\gamma$ is the unique test function that solves (IDE) on $[0, \overline{s}]$, and it is a receiver-optimal test function. It is strictly increasing in γ (and hence c) in the Blackwell informativeness order, and converges to the fully informative test function as $c \to 1$. As a consequence, the payoff of the receiver is also strictly increasing in γ. Furthermore, $\mathcal{H}^*_\gamma$ is more Blackwell informative than any other receiver-optimal test function at γ. 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 $c = 0$.*

Proof. Here, we only show that $\mathcal{H}^*_\gamma$ is indeed receiver-optimal and leave the rest of the proof for the appendix. To see why, suppose that $\mathcal{H}$ is another test function such that $\mathcal{H}(0) > \mathcal{H}^*_\gamma(0)$. Without loss of generality, we can take this function to be linear below 0, and let κ be its slope below 0. Then $\kappa > \kappa^*_\gamma$ as $\kappa \underline{s} = \mathcal{H}(0) > \mathcal{H}^*(0) = \kappa^*_\gamma \underline{s}$. Let $x' = \min\{x \in [0, \overline{s} : \mathcal{H}(x) = \mathcal{H}^*_\gamma(x)\}$ be the smallest crossing point[17] between $\mathcal{H}$ and $\mathcal{H}^*_\gamma$. Then, we must have

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

[16] Note that we can get rid of the subscript ℓ as writing that H_ℓ satisfies this equality implies that it is continuous, and therefore $H_\ell = H$.

[17] It exists as the minimum of a nonempty ($\mathcal{H}(\overline{s}) = \mathcal{H}^*_\gamma(\overline{s})$) compact (by continuity of $\mathcal{H} - \mathcal{H}^*_\gamma$) real subset.

Then

$$\begin{aligned} H_\ell(x') - \frac{x}{(\underline{s}+x)(x-\mu_0)}\mathcal{H}(x') &\leq H^*_\gamma(x') - \frac{x}{(\underline{s}+x)(x-\mu_0)}\mathcal{H}^*_\gamma(x') \\ &= \frac{\kappa^*_\gamma \underline{s}}{\underline{s}+x} - \frac{\gamma x}{(x-\mu_0)(\underline{s}+x)} \\ &< \frac{\kappa \underline{s}}{\underline{s}+x} - \frac{\gamma x}{(x-\mu_0)(\underline{s}+x)}, \end{aligned}$$

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

Hence, our optimal test uses a rich set of signals. The following proposition describes its properties further.

Proposition 6. *The signal (i.e. expectations) distribution generated by our optimal test has support on* $\{-\underline{s}\} \cup \left[0, \overline{s}\right]$*, with atoms at* $-\underline{s}$ *and* $\overline{s}$*, and a positive, continuously differentiable, and decreasing density on* $\left[0, \overline{s}\right)$*. The signal distribution generated by the high type has support on* $\left[0, \overline{s}\right]$*, with a positive, continuously differentiable, and decreasing density on* $\left[0, \overline{s}\right)$*, and a single atom at* $\overline{s}$*. The signal distribution of the low type has support on* $\{-\underline{s}\} \cup \left[0, \overline{s}\right]$*, with a single atom at* $-\underline{s}$*, and a positive, continuously differentiable, and decreasing density on* $\left[0, \overline{s}\right)$*. Furthermore, the signal distribution generated by the high type first-order stochastically dominates that of the low type.*

Our optimal test uses a continuum of signals despite the fact that types and actions are binary. The richness of optimal tests is only in the "passing" signals as only one signal is associated with failure. There is a clustering of signals close to the threshold as illustrated on FIGURE. Intuitively, enriching the set of signals that lead to approval allows the receiver to get better information while discouraging falsification. Increasing falsification would increase the probability that the low type generates the continuum of nonegative signals rather than the reject signal. But the receiver would react by rejecting some of the nonnegative signals above in an amount that exactly offsets the advantage from the first effect.

Our optimal test makes the agent indifferent across all moderate levels of falsification as it satisfies (IDE). Indifference of the "agent" at the optimal information structure also appears in Roesler and Szentes (2017) or Chassang and Ortner (2016). In our context, a test which makes no-falsification strictly better than some other falsification threshold cannot be optimal, since

it is possible to increase the informativeness of that test and still maintain that no falsification is a best response for the agent.

Both our optimal test and the three-signal test from SECTION EXAMPLE[18] deliver at least 50% of the full information payoff. A numerical analysis 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.

5.4 The General Case without Falsification Cost

6 Concluding Remarks

Here I pasted some OLD materials

We study optimal tests in the presence of falsification. Our results deliver insights for how to enhance the reliability of tests that agents can manipulate. First, fully revealing tests—albeit optimal in the absence of falsification—are prone to manipulations, and yield the worst possible results. More generally, our analysis of a binary-state, binary-action setup highlights that simple (binary) tests can be fully manipulated by the agent: Any binary test can be turned to deliver the agent-optimal information structure. Tests that perform well have more grades than actions, and must assign intermediate grades with sufficiently high probability. In fact, the simple addition of a third signal can go a long way towards optimality. We show that the optimal three-signal test delivers at least around 80% of the payoff of the optimal test, and 50% of the full information payoff. This test contains a simple practical insight: introducing a "noisy" (pooling) grade that is associated with approval in the absence of falsification, can make falsification so costly that it prevents it, rendering this noisy test much better than the (manipulated) fully informative test.

To illustrate the logic of the optimal test, consider how a four-signal approximation of our optimal test could work in practice. Such a test could have grades A, B, C, D, where A, B, C all lead to approval, but are associated with decreasingly strong beliefs about type, and D is a reject signal. In the event that manipulations are observed, grades are devalued so as to counteract the benefit of manipulations for the agent. For example, if manipulations are moderate, A, B still lead to approval, but C is devalued to a reject grade. Under greater manipulation, B or

[18] We can show that this test is in fact the optimal three-signal test.

even A and B can be devalued to reject grades as well.

Our analysis can be extended in several ways. First, falsification decisions can take place after the agent knows the types of his item(s) (interim falsification). Second, we can accommodate multiple agents, each choosing falsification rates independently of one-another. Persuaders then face a free-rider's problem, as if others do not falsify, the "penalty" each falsifier faces in terms of signal devaluation is smaller. We can account for this by modifying the non-falsification constraint.

Other interesting extensions include the possibility of adding aggregate uncertainty and endogenous priors. Suppose that receivers are uncertain about μ_0, while the agent knows the true μ_0. Then using our optimal test for a particular value μ_0' would lead each agent with a different realization μ_0 to falsify so as to generate the same grade distribution as a agent with μ_0' and no falsification. So a agent with $\mu_0 > \mu_0'$ would set $p_G > 0$, and a agent with $\mu_0 < \mu_0'$ would set $p_B > 0$. This implies that using such a test with a value μ_0' in the support of possible μ_0 would lead to small variations in performance when the support is sufficiently narrow. However, deriving the optimal test would require a different analysis. One possibility would be for the principal to design menus of tests leading different types μ_0 to self select in the spirit of Kolotilin, Li, Mylovanov, and Zapechelnyuk (2016). Such an analysis and whether menus could be useful is beyond the scope of this paper.

Suppose, now, that μ_0 is unobservable and endogenous in the sense that the fraction of good items in the market depends on how much effort the agent exerts. If production costs are sufficiently low, then the agent will set $\mu_0 \geq \hat{\mu}$ as, with such a prior, all items are approved regardless of the test, since any test can be turned to a completely uninformative one. If it is sufficiently costly to increase μ_0, then, in equilibrium, regardless of the test, only the least costly prior–say μ_L– is chosen. Otherwise, μ_L-agent can mimic the empirical distribution of grades of $\mu_H \neq \mu_L$ by falsifying as described in the previous paragraph. Hence the optimal test with moral hazard is our optimal test calibrated to $\mu_0 = \mu_L$.

Appendix

Proof of Proposition 1. Suppose that falsification is costless so $C(\phi) = 0$ and that the test is τ and the optimal falsification is ϕ, that is, for all $\tilde{\phi} : S \to \Delta(S)$:

$$U(h, \phi) \geq U(h, \tilde{\phi}). \tag{2}$$

Define a new test as follows:

$$\tau^{FP}(x|s) \equiv \int_{t \in S} \tau(x|t)\phi(t|s)dt. \tag{3}$$

With τ^{FP} and no falsification, that is $\phi(.|s) = \delta_s$, two things are true: First, the posterior after a every signal x is the same as with test τ and falsification ϕ–that is, (3) implies:

$$\tau\phi\pi_x(s) = \tau^{FP}\delta\pi_x(s) = \frac{\pi(s)\int_{t\in S}\tau(x|t)\phi(t|s)dt}{\int_{s'\in S}\int_{t\in S}\tau(x|t)\phi(t|s')dt\pi(s')ds'} = \frac{\pi(s)\tau^{FP}(x|s)}{\int_{s'\in S}\tau^{FP}(x|s')\pi(s')ds'}. \tag{4}$$

Thus, the set of "approve" signals stays the same, regardless of whether falsification is observable or unobservable:

$$\bar{X}(\tau, \phi) = \bar{X}(\tau^{FP}, \delta). \tag{5}$$

Second, the probability that each state s is approved is the same under (τ, ϕ) and (τ^{FP}, δ). Then, it follows that:

$$U(\tau, \phi) = U(\tau^{FP}, \delta) \text{ and } V(\tau, \phi) = V(\tau^{FP}, \delta). \tag{6}$$

We now need to verify that given test τ^{FP} the agent does not want to deviate from δ. We argue by contradiction. Suppose that the agent benefits by falsifying τ^{FP}, say by using falsification strategy $\phi^{'}$. The resulting posterior after a signal x is

$$\tau^{FP}\phi^{'}\pi_x(s) = \frac{\pi(s)\int_{t\in S}\tau^{FP}(x|t)\phi'(t|s)dt}{\int_{s'\in S}\int_{t\in S}\tau^{FP}(x|t)\phi'(t|s')dt\pi(s')ds'}. \tag{7}$$

In order that this to be profitable it must be the case that:

$$U(\tau^{FP}, \phi') > U(\tau^{FP}, \delta). \tag{8}$$

The agent can achieve the same distribution over posteriors conditional on each signal realization, by falsifying the original test τ with the following falsification strategy:

$$\phi''(t|s) = \int_{s' \in S} \phi(t|s')\phi'(s'|s)ds'. \tag{9}$$

First note that is a valid falsification strategy since:

$$\int_{t\in S} \phi''(t|s)dt = \int_{t\in S}\int_{s'\in S} \phi(t|s')\phi'(s'|s)ds'dt = \int_{s'\in S} \underbrace{\left(\int_{t\in S} \phi(t|s')dt\right)}_{=1} \phi'(s'|s)ds' = 1.$$

Second, note that ϕ'' results in the same distribution of posteriors about s conditional on each $x \in X$.

$$\tau^{FP}\phi^{'}\pi_x(s) = \int_{t\in S} \tau^{FP}(x|t)\phi'(t|s)dt = \int_{t\in S} \tau(x|t)\phi''(t|s)dt = \tau\phi^{''}\pi_x(s). \tag{10}$$

To see why (10) holds with ϕ'', note:

$$\int_{t\in S} \underbrace{\int_{s'\in S} \tau(x|s')\phi(s'|t)ds'}_{\tau^{FP}(x|t)} \phi'(t|s)dt = \int_{s'\in S} \tau(x|s') \int_{t\in S} \phi(s'|t)\phi'(t|s)dtds' = \int_{s'\in S} \tau(x|s')\phi''(s'|s)ds'.$$

Case 1: Unobservable falsification. In the case of unobservable falsification, we set of passing signals, does not "react" so $\bar{X}(\tau, \phi'') = \bar{X}(\tau, \phi)$ and $\bar{X}(\tau^{FP}, \phi') = \bar{X}(\tau^{FP}, \delta)$ and by construction of τ^{FP}, (5) holds. Then:

$$U(\tau, \phi'') = U(\tau^{FP}, \phi'). \tag{11}$$

Now, when falsifications costs as zero, (6), (8) and (11), imply:

$$U(\tau, \phi'') > U(\tau, \phi),$$

which contradict (2).

Case 2: Unobservable falsification To obtain a contradiction in this case as well, we need to establish that the set of passing signals stays the same. Note that (10), together with (7) guarantees that the Receiver's action after *each* x stays the same, so $\bar{X}(\tau, \phi'') = \bar{X}(\tau^{FP}, \phi')$.

□

A Uncommitted Proofs

Proof of Lemma 1. Consider a τ and let ϕ be an equilibrium falsification strategy under τ. The receiver approves whenever $\mu(x|\tau,\phi) \geq 0$ and recall that we let $\bar{X}(\tau,\phi) = \{x : \mu(x|\tau,\phi) \geq 0\}$ denote the approval set of the receiver. Define a new test such that

$$\tau'(\text{approve}|s) = \int_{\bar{X}(\tau,\phi)} \tau(dx|s).$$

Note that an equivalent formulation of τ' is $\tau'(\text{approve}|s) = a(s;\tau,\phi)$ for all $s \in S$, so by construction τ' and ϕ maintain the same interim probability of approval as the one achieved by τ and ϕ.

This test generates: (i) the same distribution of actions as a function of each state s, (ii) a given probability of approval under τ' has then the same falsification costs as τ. Point (i) is by construction and it immediately implies that $V(\tau,\phi) = V(\tau',\phi)$. Point (ii) is also easy to see, since:

$$\begin{aligned}
U(\tau,\phi) &= \int_{\bar{X}(\tau,\phi)\times S} d\tau\phi\pi - \int_{S\times S} c\, d\phi\pi \\
&= \int_S \int_{\bar{X}(\tau,\phi)} \tau(dx|t) d\phi\pi - \int_{S\times S} c\, d\phi\pi \\
&= \int_S \tau'(\text{approve}|t) d\phi\pi - \int_{S\times S} c\, d\phi\pi \\
&= U(\tau',\phi).
\end{aligned}$$

The last step to show is that there are no new falsification opportunities. Suppose there was ϕ' such that

$$\begin{gathered}
U(\tau',\phi') > U(\tau',\phi) \iff \\
\int_{S\times S} \tau'(\text{approve}|t) d\phi'\pi - \int_{S\times S} c\, d\phi'\pi > \int_{S\times S} \tau'(\text{approve}|t) d\phi\pi - \int_{S\times S} c\, d\phi\pi \iff \\
\int_{S\times S} \int_{\bar{X}(\tau,\phi)} \tau(x|t) d\phi'\pi - \int_{S\times S} c\, d\phi'\pi > \int_{S\times S} \int_{\bar{X}(\tau,\phi)} \tau(x|t) d\phi\pi - \int_{S\times S} c\, d\phi\pi \iff \\
U(\tau,\phi') > U(\tau,\phi)
\end{gathered}$$

which contradicts the fact that the original falsification strategy ϕ was part of an equilibrium

given τ. Note that given that falsification is unobservable, in the third line above we used that $\bar{X}(\tau, \phi) = \bar{X}(\tau', \phi)$. □

Proof of Proposition 2. Given the first remark, we know that constraint (RO) is redundant so the programs only differ in the falsification constraint. Let ϕ^{EO}, ϕ^{IO} denote respectively an ex-ante and an interim falsification strategy. The interim program is more constrained so, trivially: $V^{EO} \geq V^{IO}$. To prove the result we need to argue that $V^{EO} \leq V^{IO}$.

Suppose, otherwise. Then, $V^{EO} > V^{IO}$, this means that at a solution of (EP), the agent has an ex-ante optimal falsification strategy, ϕ^{EO}, that is not optimal at the interim state, so there are measurable sets of states S' such that

$$\phi^{EO}(S') \neq \phi^{IO}(S'),$$

where ϕ^{IO} satisfies (IOF).

Case 1: If all these sets have measure zero, then $V^{EO} > V^{IO}$ is not possible. To see this, modify $\phi^{EO}(S')$, to get $\hat{\phi}^{EO}(S')$ as follows:

$$\hat{\phi}^{EO} = \begin{cases} \phi^{EO} \text{ when } \phi^{EO} = \phi^{IO} \\ \phi^{IO} \text{ otherwise} \end{cases}$$

Note that V^{EO} stays the same because $\hat{\phi}^{EO}$ and ϕ^{EO} differ on measure zero sets. Moreover $\hat{\phi}^{EO}$ satisfies (IOF) by construction. Then, the test τ^{EO} together with $\hat{\phi}^{EO}$ is a solution to (IP). Contradiction.

Case 2: Suppose that there is at least one strictly positive measure set of states S' where $\phi^{EO}(S') \neq \phi^{IO}(S')$. But then consider,

$$\tilde{\phi} = \begin{cases} \phi^{EO} \text{ for } s \notin S' \\ \phi^{IO}(S') \text{ otherwise .} \end{cases}$$

Note that given that this deviation is unobservable, given the test τ^{EO}, the receiver's behavior stays the same. Given that S' has strictly positive measure, the agent's payoff strictly increases,

since so

$$\int_{S\times S}\big\{\tau(t)-c(t|s)\big\}d\tilde{\phi}\pi(t,s)>\int_{S\times S}\big\{\tau(t)-c(t|s)\big\}d\phi^{EO}\pi(t,s)$$

contradicting the fact that ϕ^{EO} satisfies (EOF).

□

Proof of Lemma 2. Consider a test and an optimal falsification strategy given the test (τ,ϕ). Let

$$t(s;\tau)\in \text{argmax}_{s'\in S}=\tau(s')-c(s'|s) \tag{12}$$

be an optimal falsification target for s (in other words, $t(s;\tau)\in\Phi(s;\tau)$). Suppose we define a new test $\tilde{\tau}$ as follows:

$$\tilde{\tau}(s)=\begin{cases}\tau(t(s;\tau))-c(t(s;\tau))|s)\ \forall s\leq 0\\ \tau(s)\text{ for }s>0\end{cases} \tag{13}$$

Note that for every $s>0$, the nominal probability of passing is higher under the new test: $\tilde{\tau}(s)\geq\tau(s)$. Also, $(\tilde{\tau},\tilde{\phi})$ yields by construction the same payoff for all $s\leq 0$ as (τ,ϕ). Hence, regardless of his type, the agent is weakly better off under $\tilde{\tau}$ and any optimal falsification strategy. We proceed to argue that

$$\tilde{\phi}(s)=\delta_s\,\mathbb{1}(s<0)+\phi(s)\,\mathbb{1}(s\geq 0)$$

is an optimal falsification strategy under $\tilde{\tau}$. First we show that $\tilde{\tau}$ is falsification-proof for the negative states. Suppose, by way contradiction, that $s\leq 0$ strictly benefits by falsifying to s', then:

$$\tilde{\tau}(s')-c(s'|s)>\tilde{\tau}(s). \tag{14}$$

Consider s' and its original target $t(s',\tau)$ given τ. We show that it must be $s'<0$ and its target $t(s',\tau)\neq s'$. Suppose otherwise, that is: either (i) $s'>0$, or (ii) $s'<0$ and $t(s',\tau)=s'$. In both theses cases it must hold that: $\tilde{\tau}(s')=\tau(s')$. But then (14) implies that:

$$\tau(s')-c(s'|s)>\tau(t(s,\tau))-c(t(s,\tau)|s)$$

contradicting (12), that is the fact that $t(s)$ is an optimal target for s given test τ. Now given that $t(s') \neq s'$, we have that

$$\tilde{\tau}(s') = \tau(t(s';\tau)) - c(t(s';\tau)|s').$$

Then (14) can be equivalently written as:

$$\tau(t(s';\tau)) - c(t(s';\tau)|s') - c(s'|s) > \tilde{\tau}(s) = \tau(t(s;\tau)) - c(t(s;\tau)|s) \geq \tau(t(s';\tau)) - c(t(s';\tau)|s)$$

where the weak inequality follows from (12), but then we obtain that

$$c(t(s';\tau)|s) > c(t(s';\tau)|s') + c(s'|s)$$

which violates the triangular inequality, (TRI). Contradiction. Now we argue optimality of $\tilde{\phi}$, for positive states. By contradiction again, suppose that there exists a positive state $s \geq 0$ and some target state $t \in \Phi(s,\tilde{\tau}) \setminus \Phi(s;\tau)$. Then t must be negative, as the payoffs from falsifying as positive states has not changed. Then,

$$\tau(t') - c(t'|t) - c(t|s) = \tilde{\tau}(t) - c(t|s) > \max_u \ \tau(u) - c(u|s) \geq \tau(t') - c(t'|s),$$

where $t' \in \Phi(t;\tau)$. By (TRI), this can only hold if $t' = t$, and hence $\tilde{\tau}(t) = \tau(t)$, but then t must also be an optimal falsification target under τ.

It remains to show that $(\tilde{\tau},\tilde{\phi})$ gives better payoffs than (τ,ϕ) to both players. We have already argued that the agent is weakly better off, regardless of her type.

Next, consider the receiver. She must benefit whenever the passing probability of negative states decreases, or that of positive states increases. This is the case when switching from (τ,ϕ) to $(\tilde{\tau},\tilde{\phi})$. Under $(\tilde{\tau},\tilde{\phi})$, a negative state $s < 0$ passes with probability $\tilde{\tau}(s) = \max_u \ \tau(u) - c(u|s)$, whereas, under (τ,ϕ), it passes with probability at most

$$\max\{\tau(t) \mid t \in \Phi(s;\tau)\} \geq \tilde{\tau}(s).$$

A positive state $s \geq 0$, by contrast, obtains the sam passing probability from from (τ,ϕ) and $(\tilde{\tau},\tilde{\phi})$, hence receiver is better-off overall. □

Proof of Lemma 3. For every s, let $\gamma(s)$ be the unique state in S^- such that $c\big(\gamma(s)|0\big) = c(s|0)$ if $c(s|0) \leq c(-\underline{s}|0)$, and $\gamma(s) = s$ otherwise. Note that for negative states, we also have $\gamma(s) = s$. This function is well defined by continuity and monotonicity of the cost function. Then define the test $\tilde{\tau}$ by

$$\tilde{\tau}(s) = \max\big\{\tau(s), \tau\big(\gamma(s)\big)\big\}.$$

Hence the nominal passing probability increases weakly for nonnegative states, and remains the same for negative states. As a consequence the agent gets a better payoff under $\tilde{\tau}$ with any incentive compatible falsification strategy. Next, we build an incentive compatible falsification strategy $\tilde{\phi}$ such that nonnegative states do not falsify as negative states, and negative states do not falsify.

Consider a negative state $t \in S^-$. If there exists a nonnegative state t' such that $\gamma(t') = t$, then, for every nonnegative state $s \in S^+$, falsifying as t' dominates falsifying as t as it yields a higher passing probability $\tilde{\tau}(t') \geq \tau(t)$ at a lower cost since $c(t'|s) < c(t'|0) = c(t|0) < c(t|s)$. If there is no such t', then it must be that $c(t|0) > c(\overline{s}|0)$. Suppose that, for some nonnegative state $s \in S^+$, falsifying as t is optimal under $\tilde{\tau}$. Then

$$\tau(t) - c\big(t|\gamma(s)\big) > \tau(t) - c(t|s) = \tilde{\tau}(t) - c(t|s) \geq \tilde{\tau}(s) \geq \tau\big(\gamma(s)\big),$$

where the first inequality is from cost monotonicity, the following equality is from the definitions of $\tilde{\tau}$, the second inequality is due to the optimality of falsifying as t for s, and the last inequality is due definition of $\tilde{\tau}$. But then, comparing the first and the last term contradicts the incentive compatibility of ϕ under τ.

Because negative states do not falsify, their probability of passing is the nominal one, and is therefore unchanged. Positive states have access to passing probabilities that were already available under τ, but are now cheaper to obtain as they can be obtained by falsifying as nonnegative states. We can therefore choose $\tilde{\phi}$ so that each nonnegative state passes at least with the same probability. As a consequence, the payoff of the receiver must be at least as high as under the initial test. □

Proof of Lemma 4. (i) being obvious, we only prove (ii) and (iii). First, test monotonicity implies that, for all states, downward falsification is dominated by truth-telling. Second, for all states in either I_U or I_L, truth-telling strictly dominates falsifiying as some other state

in the same interval. Finally, if, for some state $s \in I_U$, falsifying as some state $t \in I_M \cup I_U$ strictly dominates truth-telling, then falsifying as $\min\{t, \hat{s}\}$ must strictly dominate truth-telling for state $\check{s}$. Hence, to prove (ii), we only need to check that no state $s \in I_M$ strictly prefers falsifying as some other state $t \in I_M$ to truth-telling. If that were the case, we would have

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

a violation of (TRI). Then, to prove (iii), just note that, for every $s \in I_M$,

$$\tau_{p,\hat{s}}(\hat{s}) - c(\hat{s}|s) = p - c(\hat{s}|s) = \tau_{p,\hat{s}}(s).$$

□

Proof of Proposition 4. By Lemma 2 and Lemma 3, we can assume that the pair (τ, ϕ) is such that $\phi(s) = \delta_s$ for all $s \in S^-$, and $\operatorname{supp} \phi(s) \subseteq S^+$ for all $s \in S^+$. Let $p = \sup_{s \in S^+} \tau(s)$, which exists because $\tau(S^+)$ is bounded. For every $\varepsilon > 0$, let $S(\varepsilon) = \big\{s \in S^+ : \tau(s) \geq p - \varepsilon\big\}$, and let $\bar{S}(\varepsilon)$ be the closure of $S(\varepsilon)$. By definition of p, each $S(\varepsilon)$, and hence each $\bar{S}(\varepsilon)$, is nonempty. Furthermore, $\bar{S}(\varepsilon)$ is clearly nonincreasing in ε for the inclusion order. Therefore, by Cantor's intersection theorem, $\bar{S} = \bigcap_{\varepsilon > 0} \bar{S}(\varepsilon)$ is a nonempty compact subset of S^+. Hence, $\min \bar{S}$ is well defined. Then we let

$$\hat{s} = \min\big\{\min \bar{S}, \min\{s \in S : c(s|0) \leq p\}\big\}.$$

Now consider the test-falsification pair $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$. Lemma 4 implies that $\phi_{p,\hat{s}}$ is indeed incentive compatible under $\tau_{p,\hat{s}}$. Under $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$, each nonnegative states passes with probability p, which is at least as high as the passing probability of any state under (τ, ϕ). A negative state s passes with its nominal probability in both cases, that is $\tau(s)$ under (τ, ϕ), and $\big(p - c(\hat{s}|s)\big)^+$ under $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$. Next we show that this nominal probability is lower under $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$. If $\hat{s} \neq \min \bar{S}$, then the nominal probability is 0 for all negative states under $\tau_{p,\hat{s}}$, which proves the point. Otherwise, the definition of $\hat{s} = \min \bar{S}$ implies that there exists a sequence of nonnegative states $\{t_n\}$ that converges to $\hat{s}$ and such that the sequence $\tau(t_n)$ converges to p. Then, the sequence of falsification payoffs $\tau(t_n) - c(t_n|s)$ resulting from s falsifying as t_n converges to $p - c(\hat{s}|s)$ by continuity of the cost function. Because truth-telling is optimal for negative states under τ, we have $\tau(t_n) - c(t_n|s) \leq \tau(s)$. Going to the limit, this implies $p - c(\hat{s}|s) \leq \tau(s)$. Since

$\tau(s)$ must be nonnegative, this implies that negative states pass with lower probability under $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$. Altogether, this implies that the receiver is better off under $(\tau_{p,\hat{s}}, \phi_{p,\hat{s}})$. □

Proof of Theorem 1. By Proposition 4, we only need to show that (p^*, s^*) solves the following program

$$\max_{p,\hat{s}} \; p\int_0^{\overline{s}} s dF_\pi(s) + \int_{\check{s}(p,\hat{s})}^0 s\big\{p - c(\hat{s}|s)\big\}dF_\pi(s), \tag{15}$$
$$\text{s.t. } \check{s}(p,\hat{s}) \geq 0$$

where $\check{s}(p,\hat{s}) = \min\{s \leq \hat{s} : c(\hat{s}|s) \leq p\}$ is decreasing in p and increasing in $\hat{s}$.

Suppose first that $c(\overline{s}|0) \geq 1$. Then setting $p^* = 1$ and $s^* = \max\big\{s \in S : c(s|0) \leq 1\big\}$ implies $\check{s}(p^*, s^*) = 0$, so that all nonnegative states pass with probability 1, while all negative states pass with probability 0 which is the first-best outcome for the receiver. Since the value of (15) cannot exceed the fist-best receiver payoff, (p^*, s^*) is indeed optimal.

Hence, suppose $c(\overline{s}|0) < 1$. Fixing p, it is easy to see that increasing $\hat{s}$ has no effect on the first term of the objective in (15), but strictly increases the second term. Therefore, it is optimal to set $\hat{s}$ as high as possible. If $p \geq c(\overline{s}|0)$, this means setting $\hat{s}$ to $\overline{s}$. If $p < c(\overline{s}|0)$, then this means setting $\hat{s}$ to satisfy $\check{s}(p,\hat{s}) = 0$. But then consider replacing p by $p' = c(\overline{s}|0) > p$ and $\hat{s}'$ to $\overline{s}$. The passing probability of negative states is 0 in both cases, but the passing probability of nonnegative states goes from p to $p' > p$. This shows that the optimal value of $\hat{s}$ is $s^* = \overline{s}$, and that the optimal value for p lies between $c(\overline{s}|0)$ and 1.

These restrictions imply that $\check{s}(p,\overline{s})$ lies in $[s_*, 0]$, where $s_* = \min\{s \in S : c(\overline{s}|s) \leq 1\} < 0$ satisfies $s_* < 0$. These remaining tests are characterized by the choice of $\check{s} \in [s_*, 0]$, which then implies $p = c(\overline{s}|\check{s})$, and naturally $\hat{s} = \overline{s}$. Then we can rewrite our program as

$$\max_{\check{s}\in[s_*,0]} \; c(\overline{s}|\check{s})\int_{\check{s}}^{\overline{s}} s dF_\pi(s) - \int_{\check{s}}^0 s c(\overline{s}|s) dF_\pi(s). \tag{16}$$

Let $\Omega(\check{s})$ denote the objective function in this program. Consider $\check{s}' > \check{s} \geq s_0$. The difference $\Omega(\check{s}') - \Omega(\check{s})$ can be written as:

$$\underbrace{\big\{c(\overline{s}|\check{s}') - c(\overline{s}|\check{s})\big\}}_{<0}\underbrace{\int_{\check{s}}^{\overline{s}} s dF_\pi(s)}_{\geq 0} + \int_{\check{s}}^{\check{s}'} \underbrace{s}_{\leq 0}\underbrace{\big\{c(\overline{s}|s) - c(\overline{s}|\check{s}')\big\}}_{\geq 0} dF_\pi(s) < 0.$$

Hence, $\Omega(\check{s})$ is decreasing over $[s_0, 0]$.

Now suppose $s_0 \geq \check{s}' > \check{s}$, and write the difference $\Omega(\check{s}') - \Omega(\check{s})$ as

$$\underbrace{\left\{c(\overline{s}|\check{s}') - c(\overline{s}|\check{s})\right\}}_{<0} \underbrace{\int_{\check{s}'}^{\overline{s}} s dF_\pi(s)}_{\leq 0} + \int_{\check{s}}^{\check{s}'} \underbrace{s}_{\leq 0} \underbrace{\left\{c(\overline{s}|s) - c(\overline{s}|\check{s})\right\}}_{\leq 0} dF_\pi(s) > 0.$$

Hence, $\Omega(\check{s})$ is increasing over $[s_*, s_0]$ if $s_* < s_0$.

Therefore, the optimal choice is to set $\check{s} = \max\{s_0, s_*\}$, which leads to $p^* = c(\overline{s}|\max\{s_0, s_*\}) = \min\{c(\overline{s}|s_0), 1\}$.

The equilibrium information structure is such that all positive states pass, and would therefore lead to payoffs on the Pareto frontier in the absence of falsification costs. But because this optimal test requires some falsification by nonnegative types, the resulting payoffs must be bounded away from the Pareto frontier. □

Proof of Lemma 5. First note that continuity of any falsification-proof test is implied by (FPIC), by invoking the continuity of the cost function. Next, we define the new test τ' by

$$\hat{\tau}(s) = \underline{\tau}(s)\, \mathbb{1}_{s<0} + \overline{\tau}(s)\, \mathbb{1}_{s\geq 0},$$

where $\underline{\tau} : S^- \to [0,1]$ is the greatest nondecreasing function everywhere below τ on S^-, and $\overline{\tau} : S^+ \to [0,1]$ is the smallest nondecreasing function everywhere above τ on S^+. Because $\hat{\tau}$ delivers higher nominal passing probabilities to nonnegative states, and lower ones to negative states, it is better than τ for the receiver if it is indeed falsification proof. Furthermore, it inherits the continuity of τ and is nondecreasing by construction. As such, if it violates (FPIC) for some pair (s,t), then $s < t$ and $\hat{\tau}(s) < \hat{\tau}(t)$. Let $s' = \max\{s' \leq s : \hat{\tau}(s') = \hat{\tau}(s)\}$, and $t' = \min\{t' \leq t : \hat{\tau}(t') = \hat{\tau}(t)\}$. Then $\hat{\tau}(t') = \tau(t')$ and $\hat{\tau}(s') = \tau(s')$. Furthermore, we have

$$c(t'|s') \leq c(t|s) < \hat{\tau}(t') - \hat{\tau}(s') = \tau(t') - \tau(s'),$$

where the first inequality is due to cost monotonicity. However, this contradicts falsification proofness of τ. Hence $\hat{\tau}$ is falsification proof. Lipschitz continuity follows, as, for any pair (s,t),

$$|\hat{\tau}(t) - \hat{\tau}(s)| \leq c(t \vee s | t \wedge s) \leq K|t - s|.$$

□

Proof of Lemma 6. □

Proof of Theorem 2. □

B Committed Proofs

Proof of Lemma 7. □

Proof of Lemma 8. Normalization by the mean implies that, in the absence of falsification, the likelihood ratio informally defined by $\lambda(x) = \frac{d\overline{H}(x)}{d\underline{H}(x)}$ exists for every $x < \overline{s}$ and satisfies

$$\lambda(x) = \frac{(1-\pi_0)(x+\underline{s})}{\pi_0(\overline{s}-x)},$$

which is strictly increasing in x. Hence, with falsification, this likelihood ratio is also well defined and satisfies

$$\lambda(x,\phi) = \frac{d\overline{H}(x)}{\phi d\overline{H}(x) + (1-\phi)d\underline{H}(x)} = \frac{\lambda(x)}{\phi\lambda(x)+1-\phi},$$

which is strictly increasing in x whenever $\phi < 1$. The receiver's best response is clearly to approve whenever $\lambda(x,\phi) \geq \lambda(0)$, which implies that she uses a threshold approval strategy. Note that, for $\phi > 0$, we have

$$\lim_{x\to\overline{s}} \lambda(x,\phi) = \frac{1}{\phi},$$

implying that the threshold is $\overline{s}$, whenever $\frac{1}{\phi} \leq \lambda(0)$, that is, whenever $\phi \geq \frac{\pi_0\overline{s}}{(1-\pi_0)\underline{s}}$. Otherwise, the threshold is equal to the unique x that solves $\lambda(x,\phi) = \lambda(0)$. A little algebra then yields our formula for $\hat{x}(\phi)$, and the remaining claims are trivial. □

Proof of Corollary 3. The only part that needs additional explanations is the calculation of the agent's payoff. To see this, note that the payoff is given by

$$V\big(\tau_H, \hat{\phi}(x)\big) = 1 - (\pi_0 + (1-\pi_0)\hat{\phi}(x))\overline{H}_\ell(x) + (1-\pi_0)(1-\hat{\phi}(x))\underline{H}_\ell(x) - (1-\pi_0)c\hat{\phi}(x).$$

The rest is algebra using the formulas

$$\overline{H}_\ell(x) = \frac{1}{\mu_0 + \underline{s}}\big\{(x + \underline{s})H_\ell(x) - \mathcal{H}(x)\big\},$$

and

$$\underline{H}_\ell(x) = \frac{1}{\overline{s} - \mu_0}\big\{(\overline{s} - x)H_\ell(x) + \mathcal{H}(x)\big\},$$

as well as the identity $\mu_0 = \pi_0\overline{s} - (1 - \pi_0)\underline{s}$. □

Proof of Theorem 3. We proceed in steps. Note that optimality for the receiver is proved in the main body of the paper.

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

$$\mathcal{H}(x) = \left\{\kappa\underline{s}\left(1 + \underbrace{\int_0^x \frac{1}{(\underline{s} + y)\zeta(y)}dy}_{\chi(x)}\right) - \gamma\underbrace{\int_0^x \frac{y}{(y - \mu_0)(y + \underline{s})\zeta(y)}dy}_{\xi(x)}\right\}\zeta(x),$$

where

$$\zeta(x) = \exp\left(\int_0^x \frac{y}{(y - \mu_0)(y + \underline{s})}dy\right).$$

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

$$\log\zeta(x) = \int_0^x \frac{y}{(y - \mu_0)(y + \underline{s})}dy = \left[\frac{\mu_0}{\mu_0 + \underline{s}}\log(y - \mu_0) + \frac{\underline{s}}{\underline{s} + \mu_0}\log(y + \underline{s})\right]_0^x,$$

leading to

$$\zeta(x) = \left(\frac{x - \mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0 + \underline{s}}}\left(\frac{x + \underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0 + \underline{s}}}.$$

Next

$$\xi(x) = \left[-\exp\left(-\int_0^y \frac{z}{(z - \mu_0)(z + \underline{s})}dz\right)\right]_0^x = 1 - \frac{1}{\zeta(x)}.$$

Finally, using the closed form for ζ,

$$\begin{aligned}\chi(x) &= (-\mu_0)^{\frac{\mu_0}{\mu_0+\underline{s}}}\underline{s}^{\frac{\underline{s}}{\mu_0+\underline{s}}}\int_0^x (y-\mu_0)^{-\frac{\mu_0}{\mu_0+\underline{s}}}(y+\underline{s})^{-\frac{\underline{s}}{\mu_0+\underline{s}}-1}dy \\ &= (-\mu_0)^{\frac{\mu_0}{\mu_0+\underline{s}}}\underline{s}^{\frac{\underline{s}}{\mu_0+\underline{s}}}\left[\frac{1}{\underline{s}}\left(\frac{y-\mu_0}{y+\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}\right]_0^x \\ &= \left(\frac{-\mu_0}{\underline{s}}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x-\mu_0}{x+\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}+\frac{\mu_0}{\underline{s}}\end{aligned}$$

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

$$\begin{aligned}\kappa_\gamma^* &= \frac{\overline{s}-\mu_0}{\underline{s}\big(1+\chi(\overline{s})\big)\zeta(\overline{s})}+\gamma\frac{\zeta(\overline{s})-1}{\underline{s}\zeta(\overline{s})\big(1+\chi(\overline{s})\big)} \\ &= \kappa_0^*+\gamma\frac{\zeta(\overline{s})-1}{\underline{s}\zeta(\overline{s})\big(1+\chi(\overline{s})\big)}\end{aligned} \tag{17}$$

Step 2: $\mathcal{H}_\gamma^*$ is a test function. By construction, $\mathcal{H}_\gamma^*(\underline{s})=0$ and $\mathcal{H}_\gamma^*(\overline{s})=\overline{s}-\mu_0$. Furthermore, we see on its closed form expression that $\mathcal{H}_\gamma^*$ is twice continuously differentiable, with

$$H_\gamma^*(x)=\kappa_\gamma^*+\big(\kappa_\gamma^*(\mu_0+\underline{s})-\gamma\big)\frac{x}{(x+\underline{s})(x-\mu_0)}\left(\frac{x-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}\mathbb{1}(x>0),$$

and, differentiating once more,

$$h_\gamma^*(x)=\big(\kappa_\gamma^*(\mu_0+\underline{s})-\gamma\big)\frac{1}{(x+\underline{s})(x-\mu_0)}\left(\frac{x-\mu_0}{-\mu_0}\right)^{-\frac{\underline{s}}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{-\frac{\mu_0}{\mu_0+\underline{s}}}\mathbb{1}(x>0).$$

This density has the same sign as $\big(\kappa_\gamma^*(\mu_0+\underline{s})-\gamma\big)$ for $x>0$, implying that it is strictly positive since

$$\begin{aligned}\kappa_\gamma^*(\mu_0+\underline{s})>\gamma &\Leftrightarrow \overline{s}-\mu_0>\gamma\left(1+\frac{\overline{s}-\mu_0}{\underline{s}+\mu_0}\right)=c(\overline{s}-\mu_0) \\ &\Leftrightarrow c<1.\end{aligned}$$

Hence $\mathcal{H}_\gamma^*$ is convex and increasing. This implies in particular that it lies below the fully informative test function $\mathcal{H}_{FI}$. It remains to show that $\mathcal{H}_\gamma^*$ also lies above the uninformative

test function $\mathcal{H}_{NI}$. Here we will only show that this is true when $\gamma = 0$. We will show in step 3 below that, for every $c \in (0,1)$, $\mathcal{H}_{FI} \geq \mathcal{H}^*_\gamma \geq \mathcal{H}^*_0$ which will expand the conclusion to any γ.

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

$$\begin{aligned}
\overline{s} - \mu_0 = \mathcal{H}^*_0(\overline{s}) &= \kappa^*_0(\overline{s}+\underline{s}) - \kappa^*_0(\mu_0+\underline{s}) + \kappa^*_0(\mu_0+\underline{s}) \left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}} \left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}} \\
&= \kappa^*_0(\overline{s}-\mu_0) + \kappa^*_0(\overline{s}-\mu_0) \left(\frac{\underline{s}+\mu_0}{-\mu_0}\right) \left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{-\underline{s}}{\mu_0+\underline{s}}} \left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}} \\
&= \kappa^*_0(\overline{s}-\mu_0) \Bigg\{ 1 + \underbrace{\left(\frac{\underline{s}+\mu_0}{-\mu_0}\right) \left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{-\underline{s}}{\mu_0+\underline{s}}} \left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}}_{\geq 0} \Bigg\}
\end{aligned}$$

implying the result.

Step 3: comparative statics with respect to γ. Note that (17) implies that, for $x \in (\underline{s}, 0)$, $\mathcal{H}^*_\gamma(x) > \mathcal{H}^*_0(x)$, as it is easy to see that $\zeta(x) > 1$. Using (17), and the functions we defined in step 1, we can write that, for $x > 0$,

$$\mathcal{H}^*_\gamma(x) = \frac{\zeta(x)\big(1+\chi(x)\big)}{\zeta(\overline{s})\big(1+\chi(\overline{s})\big)}(\overline{s}-\mu_0) + \gamma\big(1+\chi(x)\big)\zeta(x) \underbrace{\left(\frac{\big(\zeta(\overline{s})-1\big)}{\zeta(\overline{s})\big(1+\chi(\overline{s})\big)} - \frac{\zeta(x)-1}{\big(1+\chi(x)\big)\zeta(x)}\right)}_{A(x)}.$$

To show that $A(x) > 0$ on $(0, \overline{s})$, we show that the function $\frac{\zeta(x)-1}{\big(1+\chi(x)\big)\zeta(x)}$ is increasing by calcu-

lating its derivative:

$$\begin{aligned}
\left(\frac{\zeta(x)-1}{\big(1+\chi(x)\big)\zeta(x)}\right)' &\propto (1+\chi(x))\zeta'(x) - \chi'(x)\zeta(x)\big(\zeta(x)-1\big) \\
&= \frac{x}{(x-\mu_0)(x+\underline{s})}\big(1+\chi(x)\big)\zeta(x) - \frac{1}{x+\overline{s}}\big(\zeta(x)-1\big) \\
&\propto \big(\mu_0 + x\chi(x)\big)\zeta(x) + (x-\mu_0) \\
&= \left\{\frac{\mu_0}{\underline{s}}(x+\underline{s}) + \left(\frac{-\mu_0}{\underline{s}}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x-\mu_0}{x+\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}\right\}\left(\frac{x-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}} \\
&\quad + (x-\mu_0) \\
&= \mu_0\left(\frac{x-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{1+\frac{\underline{s}}{\mu_0+\underline{s}}} + \frac{1+\underline{s}}{\underline{s}}(x-\mu_0) \\
&\propto -\left(\frac{x-\mu_0}{-\mu_0}\right)^{-\frac{\underline{s}}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{1+\frac{\underline{s}}{\mu_0+\underline{s}}} + \frac{1+\underline{s}}{\underline{s}} \geq \frac{1}{\underline{s}} \\
&> 0,
\end{aligned}$$

where the last inequalities are obtained by noticing that $-\left(\frac{x-\mu_0}{-\mu_0}\right)^{-\frac{\underline{s}}{\mu_0+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{1+\frac{\underline{s}}{\mu_0+\underline{s}}}$ is increasing (its derivative is proportional to $(\underline{s}-\mu_0)x + (\mu_0+\underline{s})^2$), and therefore bounded below by its value at 0 which is equal to 1.

This shows that, for every $x \in (0, \overline{s})$, $\mathcal{H}^*_\gamma(x)$ is increasing in γ and furthermore $\mathcal{H}^*_\gamma(x) > \mathcal{H}^*_0(x)$. The same holds on $(-\underline{s}, 0]$ by (17). This proves he comparative statics with respect to the Blackwell informativeness ordering. The comparative statics for the receiver's payoff also follows.

Step 4: Optimality for the receiver. This argument is the main body of the text.

Step 5: Pareto efficiency. Consider any test function $\mathcal{H}$ that delivers a fixed receiver payoff P, so $\mathcal{H}(0) = P$. To maximize the agent's payoff while giving at leat P to the receiver, one needs to minimize $H_\ell(0)$ while ensuring $\mathcal{H}(0) \geq P$. By convexity of $\mathcal{H}$, this is achieved if and only if $\mathcal{H}$ is linear between $-\underline{s}$ and 0. Therefore the set of Pareto efficient test functions is exactly the set of test functions that are linear below 0.

Step 6: Payoff bound. The full information payoff of the receiver is $\pi_0\overline{s} = \frac{\underline{s}+\mu_0}{\overline{s}+\underline{s}}\overline{s}$. First, to obtain a lower bound on the payoff ratio, note that the three-signal test we obtained in SECTION EXAMPLE yields a payoff equal to $\pi_0\overline{s}\left(\frac{\underline{s}+\overline{s}}{\underline{s}+2\overline{s}}\right) \geq \frac{1}{2}\pi_0\overline{s}$ in the absence of cost. Since

our optimal test does better, it delivers more than one half of the full information payoff in the absence of falsification cost, and yet more with a positive cost. Next, we show that the bound is tight in the absence of cost. To see this note that the payoff ratio can be written

$$\begin{aligned}\frac{\mu_0+\mathcal{H}_0^*(0)}{\frac{\underline{s}+\mu_0}{\overline{s}+\underline{s}}\overline{s}} &= \frac{\mu_0(\overline{s}+\underline{s})}{(\underline{s}+\mu_0)\overline{s}} + \frac{\kappa_0^*\underline{s}(\overline{s}+\underline{s})}{(\underline{s}+\mu_0)\overline{s}} \\ &= \frac{\mu_0(\overline{s}+\underline{s})}{(\underline{s}+\mu_0)\overline{s}} + \frac{(\overline{s}-\mu_0)\underline{s}(\overline{s}+\underline{s})}{(\underline{s}+\mu_0)\overline{s}\left(\overline{s}-\mu_0+(\underline{s}+\mu_0)\left(\frac{\overline{s}-\mu_0}{-\mu_0}\right)^{\frac{\mu_0}{\mu_0+\underline{s}}}\left(\frac{\overline{s}+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_0+\underline{s}}}\right)}.\end{aligned}$$

Choosing the parameters: $\underline{s} = 1/n + 1/n^2$, $\mu_0 = -1/n^2$ and $\overline{s} = 1 - 1/n + 1/n^2$, and replacing, we get that this ratio is equal to

$$R_n = \frac{n}{n(n-1)-1} + \frac{(n-1)(n+1)}{n^2(1-1/n-1/n^2)\left(1-1/n+\left(\frac{n^2}{n^2-1}\right)^{1/n}\left(\frac{n^2}{(n+1)n}\right)\right)},$$

which converges to $1/2$ as $n \to \infty$.

Step 7: $\mathcal{H}_\gamma^*$ is more informative than any other receiver-optimal test. First, if $\mathcal{H}$ is another receiver-optimal test, we can linearize it to the left of 0 which makes it more informative. Next, suppose that, for some $\hat{x} \in (0, \overline{s})$, $\mathcal{H}(\hat{x}) > \mathcal{H}_\gamma^*(\hat{x})$. Then, we can replicate the optimality argument of step 4 to find a contradiction. Therefore, for all $x \in (0, \overline{s})$, we have $\mathcal{H}(x) \leq \mathcal{H}_\gamma^*(x)$. Since the two test functions must be equal to the left of 0 as they are linear and deliver the same receiver payoff, we can conclude that $\mathcal{H}$ is less informative than $\mathcal{H}_\gamma^*$. □

Proof of Proposition 6. □

C Additional Proofs

Proofs for Example 1. □

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," .

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.

Ball, I. and D. Kattwinkel (2019): "Probabilistic verification in mechanism design," in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 389–390.

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

Bhaskar, D. and E. Sadler (2019): "Resource Allocation with Positive Externalities," *Available at SSRN 2853085*.

Bizzotto, J., J. Rudiger, and A. Vigier (2016): "Delegated Certification," Working paper.

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.

Boleslavsky, R. and K. Kim (2017): "Bayesian Persuasion and Moral Hazard," .

Chakravarty, S. and T. R. Kaplan (2013): "Optimal allocation without transfer payments," *Games and Economic Behavior*, 77, 1–20.

Chassang, S. and J. Ortner (2016): "Making Corruption Harder: Asymmetric Information, Collusion and Crime," Working paper.

Chua, G. A., G. Hu, and F. Liu (2019): "Optimal Multi-unit Allocation with Costly Verification," *Available at SSRN 3407031*.

Cohn, J. B., U. Rajan, and G. Strobl (2016): "Credit ratings: strategic issuer disclosure and optimal screening," Working paper.

Condorelli, D. (2012): "What money can't buy: Efficient mechanism design with costly signals." *Games Econ. Behav.*, 75, 613–624.

Condorelli, D. and B. Szentes (2016): "Buyer-Optimal Demand and Monopoly Pricing," Tech. rep., Mimeo, London School of Economics and University of Essex.

Crawford, V. P. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

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

Deneckere, R. and S. Severinov (2017): "Screening, Signalling and Costly Misrepresentation," Tech. rep., Working paper.

Epitropou, M. and R. Vohra (2019): "Optimal On-Line Allocation Rules with Verification," in *International Symposium on Algorithmic Game Theory*, Springer, 3–17.

Frankel, A. and N. Kartik (2019a): "Improving information from manipulable data," *arXiv preprint arXiv:1908.10330*.

——— (2019b): "Muddled information," *Journal of Political Economy*, 127, 1739–1776.

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 (2014): "Costly Persuasion," *American Economic Review*, 104, 457–462.

——— (2016a): "Bayesian persuasion with multiple senders and a rich signal space," .

——— (2016b): "A Rotschild-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.

Golosov, M. and A. Tsyvinski (2007): "Optimal Taxation with Endogenous Insurance Markets," *Quarterly Journal of Economics*, 122, 487–534.

Grochulski, B. (2007): "Optimal Nonlinear Income Taxation with Costly Tax Avoidance," *Economic Quarterly - Richmond Fed*.

Guo, Y. and J. Hörner (2018): "Dynamic allocation without money," *Northwestern University and Yale University*.

Guo, Y. and E. Shmaya (2018): "Costly miscalibration," Tech. rep., working paper.

Hörner, J. and N. S. Lambert (2016): "Motivational ratings," Working paper.

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. 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," .

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. (2016): "Optimal Information Disclosure: A Linear Programming Approach," Working paper.

Kolotilin, A., M. Li, T. Mylovanov, and A. Zapechelnyuk (2016): "Persuasion of a Privately Informed Receiver," Working paper.

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.

Li, R. (2020a): "Persuasion with Strategic Reporting," *Available at SSRN 3536404*.

Li, Y. (2020b): "Mechanism design with costly verification and limited punishments," *Journal of Economic Theory*, 186, 105000.

Lipman, B. (2015): "An elementary proof of the optimality of threshold mechanisms," Tech. rep., working paper, July.

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

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

Mylovanov, T. and A. Zapechelnyuk (2017): "Optimal allocation with ex post verification and limited penalties," *American Economic Review*, 107, 2666–94.

Nguyen, A. and T. Y. Tan (2019): "Bayesian Persuasion with Costly Messages," *Available at SSRN 3298275*.

——— (2020): "Bayesian Persuasion with Costly Messages," *Available at SSRN 3298275*.

Patel, R. and C. Urgun (2017): "Costly inspection and money burning in internal capital markets," *Available at SSRN 3144149*.

Perez-Richet, E. and V. Skreta (2018): "Test design under falsification," *Work. Pap., Sci. Po, UCL and UT Austin*.

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

Rochet, J. C. (1985): "The Taxation Principle and Multi-Time Hamilton-Jacobi Equations," *Journal of Mathematical Economics*, 14, 113–128.

Rodina, D. (2016): "Information Design and Career Concerns," Tech. rep., Working Paper.

Rodina, D. and J. Farragut (2016): "Inducing Effort through Grades," Tech. rep., Working paper.

Roesler, A.-K. and B. Szentes (2017): "Buyer-Optimal Learning and Monopoly Pricing," American Economic Review, forthcoming.

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," .

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

Zhang, H., Y. Cheng, and V. Conitzer (2019a): "Distinguishing Distributions When Samples Are Strategically Transformed," in *Advances in Neural Information Processing Systems 32*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Curran Associates, Inc., 3187–3195.

——— (2019b): "When Samples Are Strategically Selected," in *Proceedings of the 36th International Conference on Machine Learning*, ed. by K. Chaudhuri and R. Salakhutdinov, Long Beach, California, USA: PMLR, vol. 97 of *Proceedings of Machine Learning Research*, 7345–7353.